QUALITATIVE PROPERTIES
OF SOME DISCRETIZED PARTIAL
DIFFERENTIAL EQUATIONS AND
RELIABLE FUEL CELL MODELLING

TAMÁS SZABÓ

PhD Thesis

Supervisor
Professor ISTVÁN FARAGÓ
Doctor of the Hungarian Academy of Sciences

Mathematical Doctoral School
Director: Professor MIKLÓS LACZKOVICH
Member of the Hungarian Academy of Sciences

Doctoral Program: Applied Mathematics
Director of Program: Professor GYÖRGY MICHALETZKY
Doctor of the Hungarian Academy of Sciences

Department of Applied Analysis and Computational Mathematics
Institute of Mathematics, Faculty of Sciences
EÖTVÖS LORÁND UNIVERSITY, Budapest, Hungary

2011
I’m very thankful to Dr. Sergey Korotov (Ikerbasque Professor at the BCAM-Basque Center for Applied Mathematics) for his high-valued advices, our joint works and his hospitality I received during my first months in Bilbao, Spain.

During the last few months I have had the pleasure to meet a number of great scientists and talented students in the BCAM-Basque Center for Applied Mathematics. I am very grateful for their assistance and hospitality and I would like to thank the BCAM management for organizing excellent courses and seminars. Hereby, I acknowledge the financial support of the center.
Acknowledgements

First of all, I have a great pleasure to express my deep gratitude to my highly respected supervisor, Dr. István Faragó for his endless patient and support. I would have never started the doctoral school of applied mathematics without him and his course on mathematical analysis for meteorologists. With his help and guidance I have been able to participate in various events where I have met many great mathematicians. I was really impressed by the friendly atmosphere of the mathematical community.

Moreover, I am very grateful to Ákos Kriston (Eötvös Loránd University), who guided me into the world of fuel cells, and let me use my creativity not just in the field of science, but also for real-life problems. I thank very much to Dr. György Inzelt (Eötvös Loránd University) for his work as a scientific leader of the FuelCell.hu research group.

I extend my sincere thanks to Dr. Ágnes Havasi (Eötvös Loránd University) for our collaboration and for the language verification of my works.

I would like to express my thanks to Dr. Róbert Horváth (Budapest University of Technology and Economics) for our joint works on the parameter estimation, and to Dr. Ferenc Izsák (Eötvös Loránd University) for our work on developing numerical models.

I am thankful to the researchers, workers and former PhD fellows from the Department of Applied Analysis and Computational Mathematics at the Institute of Mathematics of Eötvös Loránd University for their support and the wonderful atmosphere they have created.

I am grateful for the support of the Hungarian National Office of Research and Technology (OMFB-00121-00123/2008), and the National Research Agency (TÁMOP-4.2.1/B-09/1/KMR-2010-003).

Last but not least, I would like to express my deepest gratitude to my family and my closest friends for their support, understanding, endless patience and (in a good case) for the festive cake.
# Table of Contents

1 Introduction ................................................................. 1  
  1.1 Preliminaries ...................................................... 1  
  1.2 Basic Definitions ................................................. 3  
  1.3 Electrochemical Terms ............................................. 7  

2 Maximum Principles for Elliptic Problems ......................... 9  
  2.1 Introduction ....................................................... 9  
  2.2 Model Problem ..................................................... 12  
  2.3 Continuous Maximum Principle ................................... 12  
  2.4 Algebraic Analogue of the Discrete Maximum Principle ...... 14  
    2.4.1 Applications for Finite Element Methods ................. 17  
    2.4.2 Applications for Finite Difference Methods .......... 20  
  2.5 Maximum Principles for Problems with Third Boundary Conditions ... 21  
    2.5.1 Applications for Finite Element Methods ................. 22  
    2.5.2 Applications for Finite Difference Methods .......... 25  
  2.6 Numerical Experiments ........................................... 30  

3 Non-negativity Preservation for Parabolic Problems ............... 35  
  3.1 Model Problem ..................................................... 36  
  3.2 Semidiscrete Solutions ........................................... 37  
    3.2.1 One-Dimensional Case ....................................... 38  
    3.2.2 Two-Dimensional Case ...................................... 39  
  3.3 Qualitative Properties of the Fully Discretized Numerical Model . 41  
    3.3.1 Non-negativity of the Iteration Matrix .................. 43  
    3.3.2 Non-negativity of Finite Element Schemes ............... 47  
    3.3.3 Tridiagonal Matrix Algorithm ............................. 53  
    3.3.4 Applications of the Tridiagonal Matrix Algorithm ...... 58  
  3.4 Numerical Experiments ........................................... 63  
  3.5 On Relations Between Qualitative Properties .................. 68
# TABLE OF CONTENTS

## 4 Proton Exchange Membrane Fuel Cells
- 4.1 History of Fuel Cells ........................................... 72
- 4.2 Modeling of Fuel Cells ........................................ 74
- 4.3 Governing Equations .......................................... 75
- 4.4 Macro-Homogeneous Model .................................. 80
- 4.5 Application of Operator Splitting ............................ 85
- 4.6 Heterogeneous Model .......................................... 89
  - 4.6.1 Numerical Solution and its Properties ................ 90
  - 4.6.2 Test Problem ................................................ 96
  - 4.6.3 Coupled Diffusion Problem ............................... 98
- 4.7 Numerical Experiments ....................................... 100

## 5 Real-Life Applications of Fuel Cells .......................... 113
- 5.1 Car-Sharing Service ........................................... 113
- 5.2 Cogeneration Power Plant .................................... 115

## Summary .................................................................. 118

## Összefoglalás ................................................................ 119

## Bibliography .......................................................... 120
Chapter 1

Introduction

1.1 Preliminaries

Generally speaking, a partial differential equation (PDE) is such an equation that contains at least one partial derivative of the unknown function, describing phenomena that depend on more than one independent variable, which distin- guishes them from the ordinary differential equations. The theory of PDEs presents an interesting theoretical subject for deeper analysis in itself. However, the most crucial motivation of the analysis is the fact that they describe many important physical, chemical and other real-life problems.

The study of PDEs was started in the 18th century by such great scientists as Euler, D’Alembert, Lagrange and Laplace. Such equations played a key role in their works as a tool for the analytical study of models appearing in physics. Mostly by the work of Riemann in the second half of the 19th century, PDEs became also very useful tools in other areas of mathematics. In spite of this the analysis of real-life and physical models still remain one of the fundamental reasons of the development of PDEs.

In the classical theory of PDEs the attention was mostly paid to the questions of existence and uniqueness of the solutions and the qualitative investigations of the solutions of PDEs started only in the mid-1950’s. The researchers tried to answer the questions: What kind of class of functions does the solution of a PDE model belong to? What kind of properties does this solution have ([41], [74], [75], [44], [89])?

The preservation of qualitative properties of different phenomena (or of PDEs) is becoming a more and more vital requirement during the construction of reliable numerical models ([13], [69], [80], [77], [106]). For phenomena that can be mathematically described by linear PDEs of elliptic and parabolic types (such as heat conduction, reaction-diffusion, pricing of options, etc.), the most important qualitative properties are the maximum-minimum principle ([91], and its special case, the non-negativity preservation (see Fig. 1.1), the maximum norm contractivity ([30], and the sign-stability ([52].

Most differential equations can only be solved by some numerical methods, hence, it is
Figure 1.1: Illustration of the preservation of the non-negativity property and the maximum principle by approximations in one dimension. The exact solution, which is positive (top), correct approximation (bottom), and two non-suitable approximations – one is having negative values (left) and one is violating the maximum principle (right). The dotted and solid red lines denote zero and the maximum value of the exact solution, respectively.

natural that we want to use such discrete models which preserve suitable equivalents of the original properties. The PDEs are often based on real-life problems, whose solutions could be able to change or improve the life quality and style of living of the people in many areas. In the course of the recent developments of many devices, the appropriate numerical models are often playing key roles.

Nowadays, the humanity is searching for suitable solutions of the main problems of the civilization. One of the most important problems is the increasing energy hunger combined by limited resources and decreasing reserves. One promising solution to this problem could be the fuel cells, i.e., such devices that convert the chemically bounded energy (e.g. hydrogen) directly into electricity. In spite of the fact that the fuel cells are more than a hundred years old inventions, their performance and design still needs to be improved. Useful tools to develop improvements are the mathematical models, which can describe the phenomena ongoing in fuel cells. The modern mathematical description of fuel cells is based on a system of time-dependent PDEs (of parabolic type) with a (usually) nonlinear source term. By solving this problem by some appropriate methods, we have an efficient tool that is able to reliably model the behavior of a fuel cell, hence, without any measurements and testing instruments, we are able to perform experiments to, e.g., increasing the efficiency of the fuel cells.
For better understanding of the phenomena in fuel cells the application of different mathematical models are crucial in many cases. Compared to the measurements, the numerical tests are not only cheaper, but they also take much shorter time to perform. However, the applied mathematical models must be reliable and should preserve the main qualitative characteristics as well.

### 1.2 Basic Definitions

In this introductory part, we give some basic definitions, notations, lemmas and technical results which will be used in the sequel.

**Definition (Laplace operator)** Symbol $\Delta^d$ denotes the second order partial differential operator of $C^2(\mathbb{R}^d)$ functions in the $d$-dimensional Euclidean space:

$$\Delta^d = \sum_{i=1}^{d} \frac{\partial^2}{\partial x_i^2}.$$  \hspace{1cm} (1.1)

**Definition (Matrix exponential)** Let $A \in \mathbb{R}^{n \times n}$ be an arbitrary quadratic matrix. Its exponential is defined as

$$\exp(A) = \sum_{k=0}^{\infty} \frac{1}{k!} A^k,$$  \hspace{1cm} (1.2)

which is always convergent, due to the relation

$$\|A^k\| \leq \|A\|^k.$$  \hspace{1cm} (1.3)

**Corollary** Let $A \in \mathbb{R}^{n \times n}$ be a diagonal matrix with the diagonal entries $a_i$ ($i = 1, \ldots, n$), then the matrix exponential of $A$ is the following diagonal matrix:

$$\exp(A) = \begin{bmatrix} e^{a_1} & 0 & 0 & \ldots & 0 \\ 0 & e^{a_2} & 0 & \ldots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \ldots & 0 & e^{a_{n-1}} & 0 \\ 0 & \ldots & 0 & 0 & e^{a_n} \end{bmatrix}.$$  \hspace{1cm} (1.4)
**Definition (Matrix non-negativity and positivity)** Let $A \in \mathbb{R}^{n \times m}$. Then $A$ is said to be non-negative (positive) if all its entries are non-negative (positive), i.e.,

$$a_{ij} \geq 0 \ (a_{ij} > 0) \text{ for all } i = 1, \ldots, n \text{ and } j = 1, \ldots, m,$$

(1.5)

and we write $A \geq 0 \ (A > 0)$. In particular, the notation $A \geq B \ (A > B)$ means that

$$a_{ij} \geq b_{ij} \ (a_{ij} > b_{ij}) \text{ for all } i = 1, \ldots, n \text{ and } j = 1, \ldots, m.$$

(1.6)

**Definition (Uniformly tridiagonal matrix)** The square matrix of the form

$$A = \begin{bmatrix} c & b & 0 & \cdots & 0 \\ a & c & b & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & a & c & b \\ 0 & \cdots & 0 & a & c \end{bmatrix},$$

(1.7)

where $a, b, c \in \mathbb{R}$, is called uniformly tridiagonal matrix and it is denoted as $\text{tridiag}(a, c, b)$. The matrix $A = \text{tridiag}(a, c, a)$ is called symmetric uniformly tridiagonal matrix.

**Definition (Kronecker product)** Let $A \in \mathbb{R}^{n \times m}$ and let $B \in \mathbb{R}^{p \times q}$. Then the Kronecker product $A \otimes B$ is the following $np \times mq$ block matrix:

$$A \otimes B = \begin{bmatrix} a_{11}B & \ldots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{n1}B & \ldots & a_{nn}B \end{bmatrix},$$

(1.8)

where $a_{ij}$ denote the entries of $A$.

**Definition (Kronecker sum)** Let $A \in \mathbb{R}^{n \times n}$ and $B \in \mathbb{R}^{m \times m}$. Then the Kronecker sum $A \oplus B$ is the following $nm \times nm$ matrix:

$$A \oplus B = A \otimes I_m + I_n \otimes B,$$

(1.9)

where $I_n$ and $I_m$ denote the $n \times n$ and $m \times m$ identity matrices, respectively.
**Definition** (One-pair matrix) A ∈ ℝ^{n×n} with entries a_{ij} is to be said to a one-pair matrix if
\[ a_{i,j} = \begin{cases} p_i q_j, & \text{if } i \leq j \\ p_j q_i, & \text{if } j \leq i \end{cases} \] (1.10)
where p_i and q_i are real vectors from ℝ^n.

**Definition** (Monotone matrix) Let A ∈ ℝ^{n×n}. Then it is said to be monotone if
\[ Ax \geq 0 \implies x \geq 0, \] (1.11)
for any vector x ∈ ℝ^n.

**Corollary**[10] A ∈ ℝ^{n×n} is a monotone matrix ⇔ A^{-1} exists and A^{-1} ≥ 0.

**Definition** (M-matrix) The matrix A ∈ ℝ^{n×n} with the entries a_{ij} is said to be an M-matrix [100] if
\[ a_{ij} \leq 0 \quad \text{for all } i \neq j \] (1.12)
and it is monotone.

**Corollary**[3] For a monotone matrix A there exists a vector g ∈ ℝ^n such that g > 0 for which
\[ Ag > 0. \] (1.13)

**Definition** (Infinity norm of a matrix) Let A ∈ ℝ^{n×n}. Then the infinity norm ||·||_∞ of the matrix A (with entries a_{ij}) is defined as
\[ ||A||_\infty := \max_{i=1,\ldots,n} \sum_{j=1}^{n} |a_{ij}|. \] (1.14)

**Definition** (Strictly row diagonally dominant matrix) Let A ∈ ℝ^{n×n}. Then A is said to be strictly row diagonally dominant (or SDD in short) if the values
\[ \alpha_i(A) := |a_{ii}| - r_i > 0 \quad \text{for all } i = 1,\ldots,n, \] (1.15)
where r_i is the sum of absolute values of all off-diagonal entries in the i-th row of A, i.e.,
\[ r_i := \sum_{j=1, j \neq i}^{n} |a_{ij}|. \] (1.16)
Definition (Reducible-irreducible matrix) Let $A \in \mathbb{R}^{n \times n}$, where $n \geq 2$, then it is said to be reducible if there exists a permutation matrix $P \in \mathbb{R}^{n \times n}$ such that

$$PAP^T = \begin{bmatrix} A_{1,1} & A_{1,2} \\ 0 & A_{2,2} \end{bmatrix},$$

(1.17)

where $A_{1,1} \in \mathbb{R}^{r \times r}$, $A_{2,2} \in \mathbb{R}^{(n-r) \times (n-r)}$, and $A_{1,2} \in \mathbb{R}^{r \times (n-r)}$. If no such permutation matrix exists, then $A$ is said to be irreducible. In the case of $n = 1$, $A$ is irreducible if its single entry is non-zero, and reducible otherwise.

Definition ($L^p(\Omega)$ Lebesgue spaces) Let $f$ be a real function defined in the given domain $\Omega$ with a Lipschitz continuous boundary $\partial \Omega$ that is Lebesgue measurable. Then $\int_{\Omega} f(x) dx$ denotes the Lebesgue integral of $f$. Moreover, we define the Lebesgue spaces

$$L^p(\Omega) := \{ f(x) \mid \|f\|_{L^p(\Omega)} < \infty \},$$

(1.18)

where

$$\|f\|_{L^p(\Omega)} := \left( \int_{\Omega} |f|^p(x) dx \right)^{1/p} \text{ for } 1 \leq p < \infty.$$

(1.19)

Definition In the sequel $\Omega \subset \mathbb{R}^d$ always denotes a bounded (Chapter 2) and later rectangular (Chapter 3) $d$-dimensional domain, $\partial \Omega$ and $\Gamma_D$ denotes its complete boundary and some part of its boundary, respectively. Further, $H^1_D(\Omega)$ denotes the subspace of Sobolev space $H^1(\Omega) := W^{1,2}(\Omega)$:

$$H^1_D(\Omega) := \{ v \in H^1(\Omega) \mid v|_{\Gamma_D} = 0 \}.$$

(1.20)

Definition (Finite element mesh) In this thesis, we consider the finite element mesh as a set

$$\mathcal{T}_h := \{ K_i \mid i = 1, \ldots, n \}$$

(1.21)

of elements $K_i \subset \overline{\Omega}$. These elements are usually considered as closed sets with a nonzero measure \cite{LS}, moreover $\bigcup_{i=1}^n K_i = \overline{\Omega}$ and the measure of $K_i \cup K_j$ is zero for all $i, j = 1, \ldots, n$, where $i \neq j$.

Lemma 1.2.1 (Lax-Milgram lemma) Assume that $H$ is a Hilbert space and let $a : H \times H \rightarrow \mathbb{R}$ be a bounded and coercive bilinear functional and $\phi \in H^*$. Then there exists a
uniquely determined $y \in H$ such that

$$a(x, y) = \phi(x) \quad \text{for all} \quad x \in H.$$  \hfill (1.22)

**Lemma 1.2.2** For any $\ell_1, \ell_2 \in \mathbb{R}$ and positive $\mu_1, \mu_2$ the estimates

$$\min \left\{ 0, \frac{\ell_1}{\mu_1}, \frac{\ell_2}{\mu_2} \right\} \leq \frac{\ell_1 + \ell_2}{\mu_1 + \mu_2} \leq \max \left\{ 0, \frac{\ell_1}{\mu_1}, \frac{\ell_2}{\mu_2} \right\}$$  \hfill (1.23)

are valid.

**Proof:** It follows from a straightforward calculation. $\blacksquare$

### 1.3 Electrochemical Terms

In this part, some electrochemical concepts which we are going to use in Chapter 4 are presented [6].

- **The Na\textsuperscript{f}ion** (registered trademark of the American company E. I. DuPont de Nemours & Co.) is a copolymer consisting of sulfonate-terminated perfluorovinyl ether groups on a tetrafluoroethylene (Teflon) backbone. It represents one of the prototypes of modern polyelectrolytes (ionomers) with a high proton conductivity.

- In general, **double layer** of charges exists at the interface between two conducting media: One side carries a positive excess charge, which is balanced by a negative excess of equal magnitude on the other side.

- The **faradaic current** is the current that flows through the external circuit connecting the electrodes of an electrochemical cell and generated by the reduction or oxidation of some chemical substance at an electrode.

- The **open circuit potential** is the difference of electrical potential between the two electrodes of a device when there is no external load connected, i.e., when the circuit is not closed.

- **Limiting current** or Potential-independent current is achieved when the electrode process is occurring at the maximum rate possible for a given set of mass transfer conditions. The current reaches a constant (steady-state) value, which is limited only by mass transport or the rate of the chemical reaction preceding the charge transfer step.
When an electrode reaction is in equilibrium, the reaction rate in the anodic direction is equal to that in the cathodic direction. Even though the net current is zero at equilibrium, we still envisage that there is the anodic current component balanced with the cathodic one. This current value is called the exchange current. The corresponding value of current density is called the exchange current density.
Chapter 2

Maximum Principles for Elliptic Problems

2.1 Introduction

The maximum principle is an important feature of scalar second order elliptic and parabolic equations, which distinguishes them from higher order equations and systems of equations ([44], [105]). The principle, in its simplest form, was first discovered for harmonic functions: any nonconstant harmonic function \( u \) (i.e., \( \Delta u = 0 \)) assumes its minimum and maximum values only on the boundary \( \partial \Omega \) of any bounded domain \( \Omega \) in which \( u \in C(\overline{\Omega}) \),

\[
\min_{s \in \partial \Omega} u(s) < u(x) < \max_{s \in \partial \Omega} u(s) \quad \text{for all} \quad x \in \Omega.
\] (2.1)

The relation (2.1) gives, in fact, an a priori estimate for \( u(x) \) in \( \Omega \) via its values on \( \partial \Omega \). Later, (continuous) maximum principles were formulated for various second order boundary value problems (see, e.g., [44], [75], [78]). For the convenience of presentation we introduce the following simple boundary value problem of elliptic type

\[
\begin{align*}
-\Delta u + cu &= f \quad \text{in} \quad \Omega, \\
    u &= g \quad \text{on} \quad \partial \Omega,
\end{align*}
\] (2.2) (2.3)

where the constant coefficient \( c \geq 0 \), \( f \) and \( g \) are given functions.

The paper [101] by Varga in 1966 was probably the first publication devoted to the construction of discrete analogues of maximum principles, usually called discrete maximum principles. In short, that work deals with the case \( f \equiv 0 \) in (2.2)–(2.3), and analyses continuous and discrete maximum principles in the following forms:

\[
\begin{align*}
\max_{x \in \Omega} |u(x)| &\leq \max_{s \in \partial \Omega} |g(s)|, \\
\max_{i=1,\ldots,n} |u_i| &\leq \max_{j=1,\ldots,n^}\ |g_j|,
\end{align*}
\] (2.4) (2.5)

9
where $u_i$ are values of the finite difference solution at interior nodes, $g_j$ are values of $g$ at boundary nodes, moreover $n$ and $n^\partial$ denotes the number of nodes and boundary nodes, respectively. Sufficient conditions for the validity of the above discrete maximum principle were given in [101] in terms of the matrix appearing in the finite difference discretization. However, for problem (2.2)–(2.3), the corresponding continuous maximum principle, in fact, takes a more sophisticated form [75]:

$$\max_{x \in \Omega} u(x) \leq \max \{0, \max_{s \in \partial \Omega} g(s)\}, \text{ or } \max_{x \in \Omega} u(x) = \max_{s \in \partial \Omega} g(s) \text{ if } c = 0, \quad (2.6)$$

provided the sign-condition $f \leq 0$ holds. Therefore, later, in works by Ciarlet [16], Ciarlet and Raviart [17] in the 70’s, a more suitable form of discrete maximum principle adopted to the maximum principle (2.6) was proposed for finite difference and finite element approximations. In particular, for linear simplicial, block, and prismatic finite elements it reads

$$\max_{x \in \Omega} u_h(x) \leq \max \{0, \max_{s \in \partial \Omega} g_h(s)\}, \text{ or } \max_{x \in \Omega} u_h(x) = \max_{s \in \partial \Omega} g_h(s) \text{ if } c = 0, \quad (2.7)$$

where $u_h$ is a finite element solution and $g_h$ is an approximation of $g$. In both works [16], [17] several sets of sufficient conditions providing the validity of discrete maximum principle (2.7) were given. In particular, in [17] simplicial meshes and piecewise linear continuous finite element approximations were used, and for the first time sufficient geometric conditions of nonobtuseness or acuteness of triangular elements (depending on the coefficient $c$) were obtained.

Later, various generalizations of the above mentioned results were done. Thus, Lorenz [79] in 1977, Höhn and Mittelmann [51] in 1981 made attempts to derive similar geometrical conditions under which relevant discrete maximum principles hold for approximations obtained with the help of higher order finite elements. Unfortunately, even for the simplest case ($c \equiv 0$ and $f \equiv 0$ in (2.2)–(2.3)), their (sufficient) conditions on the triangular meshes turned out to be very stringent (only right or equilateral triangles are allowed) and, thus, hardly employed in real computations. Some positive results in this direction have been recently obtained by Šolín and Vejchodský (see, e.g., [90], [104], [103]) for the one-dimensional case.

Further, Christie and Hall [15] in 1984 considered the case of bilinear finite element approximations for problem (2.2)–(2.3) with $c \equiv 0$ and $f \equiv 0$. In fact, the notion of non-narrow rectangular element was introduced there as a sufficient geometric condition for the corresponding discrete maximum principle to hold (see also more recent works [59] and [64] in this respect). The case of discrete maximum principle for prismatic finite
elements has been recently analyzed in [48].

The next efforts in the analysis of discrete maximum principles were done by Krížek and Qun Lin [70] in 1995. For \( f \leq 0 \) and a sufficiently smooth function \( b(x, u, \nabla u) \) they considered the following 3D nonlinear elliptic problem:

\[
- \text{div} \left( b(x, u, \nabla u) \nabla u \right) = f \text{ in } \Omega \subset \mathbb{R}^3 \text{ and } u = 0 \text{ on } \partial \Omega, \tag{2.8}
\]

for which the corresponding continuous and discrete maximum principles take the form \( u \leq 0 \) and \( u_h \leq 0 \) (for linear elements), respectively. In addition, the effect of quadrature rules was analyzed there and discrete maximum was proved under the condition of nonobtuseness of the tetrahedral meshes used.

In all the above mentioned papers, only the cases of linear problems (besides [70]) with pure Dirichlet boundary conditions were analyzed. Then, Karátson and Korotov [57], [58] in 2005 and 2006, respectively, considered a more general case of second-order nonlinear elliptic problems with mixed boundary conditions in arbitrary space dimension. They formulated and proved the corresponding continuous and discrete maximum principles (also taking into account the effect of numerical integration).

Other works devoted to various aspects of discrete maximum principles and related issues include [13], [14], [23], [71], [72], [87], [88]. Several examples of real-life problems for which the validity of discrete maximum principles is essential are given, e.g., in papers [57] and [77]. Some results obtained for a single equation were later generalized to a system of elliptic equations in [60].

Finally, let us point out that if the discrete maximum principle is not valid, then some pathological nonphysical situations may appear. For instance, the numerical heat could flow from colder parts of the body to hotter parts [72].

In this chapter, some new modifications of the continuous and discrete maximum principles first presented by the author in [36], [37], and [38] are formulated in the form of a two-sided estimation without any sign condition on the right-hand side function \( f \). In addition, we will analyze these principle for linear elliptic problems with the third (Robin) boundary condition (which was never discussed in the literature in this respect).
### 2.2 Model Problem

Let $\Omega$ be a bounded domain, with Lipschitz continuous boundary $\partial \Omega$, then we are looking for a function $u \in C^2(\overline{\Omega})$ such that

\begin{align}
-\Delta u + cu &= f \quad \text{in } \Omega, \quad (2.9) \\
u &= 0 \quad \text{on } \partial \Omega, \quad (2.10)
\end{align}

where $c(x)$ is the reactive coefficient (non-negative) for all $x \in \Omega$, moreover $c(x), f(x) \in C(\overline{\Omega})$.

The classical solution of the above problem is known to satisfy the so-called maximum principle, which can be rewritten as follows

\[ f(x) \geq 0 \quad \text{for all } x \in \overline{\Omega} \implies \min_{x \in \overline{\Omega}} u(x) \geq 0. \quad (2.11) \]

From (2.11), however, one can only get information about the sign of the unknown function $u$ on the domain $\Omega$, which can be also often important to obtain. At the same time, there are results for various a priori (upper and lower) estimates on the magnitude of the solutions of some elliptic problems [74].

The main aim of this chapter is to combine several available theoretical estimates in order to obtain a priori two-sided bounds for the classical solutions of elliptic problems (2.9)–(2.10) with positive reactive terms for arbitrary source functions and show how to prove the validity of their discrete analogues for some well-known numerical techniques, e.g., finite difference method or finite element method.

### 2.3 Continuous Maximum Principle

In the sequel, problem (2.9)–(2.10) is considered. The key result of our work for this problem is as follows.

**Theorem 2.3.1** Let $c(x)$ and $f(x)$ in (2.9) be from $C(\overline{\Omega})$, moreover, additionally let

\[ c(x) \geq c_0 > 0 \quad \text{for all } x \in \overline{\Omega}, \quad (2.12) \]

then the following (a priori) two-sided estimates are valid for the classical solution of
problem (2.9) – (2.10):

\[
\min \left\{ 0, \min_{\xi \in \Omega} \frac{f(\xi)}{c(\xi)} \right\} \leq u(x) \leq \max \left\{ 0, \max_{\xi \in \Omega} \frac{f(\xi)}{c(\xi)} \right\}, \quad \text{for any } x \in \Omega. \tag{2.13}
\]

**Proof:** The upper estimate of \( u \) is clearly valid if \( u \leq 0 \) everywhere in \( \Omega \), i.e., when the unknown function attains its maximum on the boundary \( \partial \Omega \). Moreover, if \( u \) attains its positive maximum at some interior point \( x_0 \in \Omega \), then for the first order partial derivatives

\[
\frac{\partial u}{\partial x_i}(x_0) = 0, \quad \text{for all } i = 1, \ldots, d, \tag{2.14}
\]

and for the second order partial derivatives

\[
\frac{\partial^2 u}{\partial x_i^2}(x_0) \leq 0, \quad \text{for all } i = 1, \ldots, d \tag{2.15}
\]

hold, therefore from (2.9) and (2.12) we get

\[
u(x_0) \leq \frac{f(x_0)}{c(x_0)}, \tag{2.16} \]

from which (2.13) follows immediately. The lower estimate in (2.13) can be provided in the same way.

\[\blacksquare\]

**Remark 2.3.1** Nonhomogeneous Dirichlet boundary conditions can be treated similarly [74], i.e., the following estimates hold:

\[
\min \left\{ \min_{\xi \in \partial \Omega} u(\xi), \min_{\xi \in \Omega} \frac{f(\xi)}{c(\xi)} \right\} \leq u(x) \leq \max \left\{ \max_{\xi \in \partial \Omega} u(\xi), \max_{\xi \in \Omega} \frac{f(\xi)}{c(\xi)} \right\}, \quad x \in \Omega. \tag{2.17}
\]

*It is worth emphasizing – with respect – that the very first published paper which was purely devoted to discrete maximum principles [101] is considering the case of arbitrary Dirichlet boundary conditions, but it does not analyze another important case with nonzero source functions.*

**Remark 2.3.2** From the estimation (2.13) it is easy to derive the following important implication:

\[
f(x) \geq 0, \quad x \in \Omega \Rightarrow 0 \leq u(x) \leq \max_{\xi \in \Omega} \frac{f(\xi)}{c(\xi)}, \tag{2.18}
\]
which is a sharper (two-sided) estimation for the unknown function $u$ than the standard maximum principle \((2.11)\) guaranteeing only the sign of $u$. In what follows, we refer to \((2.13)\) as the modified maximum principle as it makes both sharpening and also generalizing of the standard maximum principle \((2.11)\).

**Remark 2.3.3** Discrete maximum principles have been widely used for proving stability and finding the rate of convergence for finite difference methods (see, e.g., \([3], [10]\) and finding the convergence of finite element approximations in the maximum norm (see, e.g., \([3], [17]\)).

### 2.4 Algebraic Analogue of the Discrete Maximum Principle

After discretization of \((2.9)\) by, e.g., such popular numerical techniques as some finite element or finite difference method we arrive at the problem of solving an $n \times n$ system of linear algebraic equations

$$
Au = F, \quad (2.19)
$$

where the vector of unknowns $u = [u_1, \ldots, u_n]^T$ approximates the unknown solution $u$ at certain selected points $x_1, \ldots, x_n$ of the solution domain $\Omega$, and the vector $F = [F_1, \ldots, F_n]^T$ approximates (in the sense related to the nature of a concrete numerical method used, see Section 2.4.1 for more details on this) the values $f(x_i)$, $i = 1, \ldots, n$ and $A \in \mathbb{R}^{n \times n}$.

In the sequel, the entries of matrix $A$ will be denoted by $a_{ij}$, and all matrix and vector inequalities appearing in the text are always understood component-wise (see \((1.6)\)). Further, if one provides that $A$ in \((2.19)\) be monotone, then $A^{-1} \geq 0$ and using the assumption that $F \geq 0$ (usually trivially can be guaranteed by $f \geq 0$ from the maximum principle \((2.11)\), e.g., for linear finite element and finite difference methods) we immediately get that $u = A^{-1}F \geq 0$. These arguments describe a standard scheme for proving the following discrete maximum principle:

$$
F \geq 0 \implies u \geq 0, \quad (2.20)
$$

which naturally implies the maximum principle \((2.11)\) \(([14], [16], [57], [70])\).
Remark 2.4.1 If we can provide more information on the entries of \( A^{-1} \) (besides \( A^{-1} \geq 0 \)), then we can estimate the vector \( u \) (e.g., not only signs of its entries, but also their magnitudes, etc.) more precisely. For example, in [100] it is shown that when \( A \) is irreducibly diagonally dominant with positive diagonal entries and nonpositive off-diagonal entries, or irreducible Stieltjes matrix, (which often happens after discretizing problem (2.9) - (2.10) by various numerical techniques), then we have an even stronger result \( A^{-1} > 0 \), which can be useful to get a better estimation on the behavior of the numerical approximations. However, the property of irreducibility is not so easy to guarantee, e.g., for finite element approximations, see [47] for several examples on that. Even sharper estimates for the entries of \( u \) can sometimes be derived if we use formulae for the exact computing \( A^{-1} \), which are available, e.g., for tridiagonal matrices appearing in the numerical solution of certain one-dimensional problems, see [85].

The following result, see [1] and [99] for its proof, is useful for the purposes of this work.

**Lemma 2.4.1** Let the square \( n \times n \) matrix \( M \) be SDD (see (1.15)). Then

\[
\| M^{-1} \|_{\infty} \leq \frac{1}{\min_{i=1,\ldots,n} \alpha_i(M)}. \tag{2.21}
\]

**Remark 2.4.2** It is worth emphasizing that diffusion-reaction problems with nonzero reaction terms often lead to SDD matrices in system (2.19) (see Section 2.4.1), which gives a reasonable chance that we can often prove suitable discrete analogues of the estimates in (2.13) for such a type of problems.

**Theorem 2.4.1** [36] Let the matrix \( A \) in system (2.19) be SDD and monotone. Then using the notations of (1.15) the following two-sided estimates for the entries of the unknown function \( u \) are valid:

\[
\min \left\{ 0, \min_{j=1,\ldots,n} \frac{F_j}{\alpha_j(A)} \right\} \leq u_i \leq \max \left\{ 0, \max_{j=1,\ldots,n} \frac{F_j}{\alpha_j(A)} \right\}, \quad i = 1, \ldots, n. \tag{2.22}
\]

**Proof:** First of all it is easy to get \( \alpha_i(A) > 0 \) for all \( i = 1, \ldots, n \), hence, \( A \) is SDD. Moreover, it is clear that the solution \( u \) of system (2.19) is the solution of the following system as well:

\[
\tilde{A}u = \tilde{F}, \tag{2.23}
\]
where \( \bar{\mathbf{A}} = \mathbf{D} \mathbf{A} \) and \( \bar{\mathbf{F}} = \mathbf{D} \mathbf{F} \). Here \( \mathbf{D} \) is a diagonal matrix with strictly positive numbers \( 1/\alpha_i(\mathbf{A}) \), \( i = 1, \ldots, n \) in its diagonal. Obviously, \( \bar{\mathbf{A}} \) is also SDD with \( \alpha_i(\bar{\mathbf{A}}) = 1 \) for all \( i = 1, \ldots, n \). In addition, \( \bar{\mathbf{A}} \) is monotone as \( \mathbf{A} \) is monotone.

Let \( \mathbf{G} := \bar{\mathbf{A}}^{-1} \) and its nonnegative entries denoted by \( g_{ij} \). As \( \mathbf{u} = \bar{\mathbf{A}}^{-1} \bar{\mathbf{F}} \), we observe that

\[
\| \mathbf{G} \|_\infty \min \left\{ 0, \min_{j=1,\ldots,n} \frac{F_j}{\alpha_j(\bar{\mathbf{A}})} \right\} \leq u_i = \sum_{j=1}^{n} g_{ij} \frac{F_j}{\alpha_j(\bar{\mathbf{A}})} \leq \| \mathbf{G} \|_\infty \max \left\{ 0, \max_{j=1,\ldots,n} \frac{F_j}{\alpha_j(\bar{\mathbf{A}})} \right\}.
\]

Now, applying Lemma 2.4.1 to \( \bar{\mathbf{A}} \), for which \( \alpha_i(\bar{\mathbf{A}}) = 1 \), we see that \( \| \mathbf{G} \|_\infty = \| \bar{\mathbf{A}}^{-1} \|_\infty \leq 1 \).

From this and the above inequalities, we finally get the required estimates (2.22).

\[\square\]

**Remark 2.4.3** It is clear that estimates (2.22) immediately imply the discrete maximum principle (2.20) provided by \( \mathbf{F} \geq 0 \). Moreover, the estimation (2.22) is considerably sharper, than the obvious, but very rough bounds \( |u_i| \leq \| \mathbf{A}^{-1} \|_\infty \| \mathbf{F} \|_\infty \).

**Remark 2.4.4** Estimates close to (2.22) were obtained earlier by Windisch in [106] (however, in a more complicated way), but only for a more restrictive case of strictly row diagonally dominant \( \mathcal{M} \)-matrices. We also notice that we could easily get an even sharper estimation in the proof of the above theorem by dropping zeros in (2.22). However, as we link the results to the continuous case, i.e., to (2.13), containing zeros, it is not actually necessary to do so in what follows.

**Remark 2.4.5** In the work by Smelov [88], a very general case of discrete maximum principle with an arbitrary SDD matrix \( \mathbf{A} \) has been considered and two-sided estimation similar to (2.22) has also been presented. However, adding a quite natural (and rather standard) requirement of monotonicity for matrices appearing in (2.19) (as we do in this work) leads to a sharper estimation (2.22), which, moreover imitates its continuous counterpart (2.13).

As (2.22) actually resembles the estimates (2.13), it is natural to give the following definition.

**Definition** We say that the solution \( \mathbf{u} \) of system (2.19) with SDD matrix \( \mathbf{A} \) satisfies the modified discrete maximum principle (or MDMP, in short), corresponding to the modified maximum principle (2.13), if estimates (2.22) are valid and if, in addition,

\[
\max_{j=1,\ldots,n} \frac{F_j}{\alpha_j(\mathbf{A})} \leq \max \left\{ 0, \max_{\mathbf{x} \in \Omega} \frac{f(\mathbf{x})}{c(\mathbf{x})} \right\}, \tag{2.24}
\]

\[
\min_{j=1,\ldots,n} \frac{F_j}{\alpha_j(\mathbf{A})} \geq \min \left\{ 0, \min_{\mathbf{x} \in \Omega} \frac{f(\mathbf{x})}{c(\mathbf{x})} \right\}. \tag{2.25}
\]
Remark 2.4.6 The conditions \((2.24)\) and \((2.25)\) are really important in order to produce reliable (i.e., controllable) numerical approximations as, for example, linear finite difference and finite element approximations do stay within the same (a priori known) limits as those of the exact solutions they do approximate.

Remark 2.4.7 While the SDD-property of \(A\) is almost automatically guaranteed after discretization by the nature of the reaction-diffusion, heat conduction, etc. problems, its monotonicity, required in Theorem \([2.4.1]\) should be provided a priori (or proved separately in each concrete case). One common approach for this in finite element methods is to impose certain a priori geometric requirements on the finite element meshes applied so that all the off-diagonal entries \(a_{ij} \leq 0\) (see, e.g., \([14]\), \([17]\), \([47]\), \([57]\), \([70]\), \([104]\) for more details on this subject). As far as it concerns finite difference method, this property for the off-diagonal entries of \(A\) is often guaranteed a priori by many standard finite difference schemes producing the so-called \(M\)-matrices \([40]\).

Remark 2.4.8 The bounds \((2.22)\) are achievable, e.g., when \(A\) is the identity matrix.

2.4.1 Applications for Finite Element Methods

In what follows we demonstrate how theoretical results of the previous sections can be used for proving modified discrete maximum principles for several popular numerical schemes (of finite element method and finite difference methods), thus increasing the level of reliability of practical calculations by these techniques.

The standard finite element methods are based on the so-called variational formulation of \((2.9)\) and \((2.10)\), which reads as follows. Find the unknown \(u \in H^1_0(\Omega)\) (see \((1.20)\), with \(\Gamma_D := \partial \Omega\)) such that
\[
a(u, v) = F(v) \text{ for all } v \in H^1_0(\Omega),
\]
where
\[
a(u, v) = \int_\Omega \nabla u \cdot \nabla v d\mathbf{x} + \int_\Omega c uv d\mathbf{x} \quad \text{and} \quad F(v) = \int_\Omega f v d\mathbf{x}.
\]

The existence and uniqueness of the (weak) solution \(u\) is provided by the standard Lax-Milgram lemma (see \((1.22)\)). It is worth emphasizing, that for the well-posedeness of the problem above one can only require that \(c \in L^\infty(\Omega)\) and \(f \in L^2(\Omega)\) (see \((1.18)\)), but we will need more smoothness from these functions in what follows.
Let $T_h$ be a finite element mesh (see (1.21)) of $\Omega$ with interior nodes $x_1, \ldots, x_n$ lying in $\Omega$ and boundary nodes $x_{n+1}, \ldots, x_{n+n^\partial}$ lying on $\partial\Omega$.

Moreover, let the basis functions $\phi_1, \phi_2, \ldots, \phi_{n+n^\partial}$, associated with these nodes, have the following properties (easily met if, e.g., simplicial or block finite element meshes are used):

$$\phi_i(x_j) = \delta_{ij}, \ i, j = 1, \ldots, n+n^\partial,$$

$$\phi_i \geq 0 \text{ in } \Omega, \ i = 1, \ldots, n+n^\partial,$$

$$\sum_{i=1}^{n+n^\partial} \phi_i \equiv 1 \text{ in } \Omega,$$

where $\delta_{ij}$ is the Kronecker delta. We also assume that the basis functions $\phi_1, \phi_2, \ldots, \phi_n$ vanish on the boundary $\partial\Omega$, thus spanning a finite-dimensional subspace denoted by $V^0_h$ of $H^1_0(\Omega)$.

The finite element approximation of $u$ is defined as a function $u_h \in V^0_h$ such that

$$a(u_h, v_h) = F(v_h) \text{ for all } v_h \in V^0_h,$$

whose existence and uniqueness are also provided by the Lax-Milgram lemma.

$u_h = \sum_{i=1}^n u_i \phi_i$, where the coefficients $u_i$ are the entries of the solution $u$ of system (2.19) with $a_{ij} = a(\phi_i, \phi_j)$ and $F_i = F(\phi_i)$. It is clear that, if (2.28)–(2.30) hold, the finite element approximation $u_h$ satisfies the bounds from (2.22) at each point of $\Omega$ if all values $u_i$ do satisfy them.

Moreover, the diagonal entries $a_{ii} = a(\phi_i, \phi_i) > 0$, $i = 1, \ldots, n$. Assume that all the applied finite element meshes used are such that $a_{ij} \leq 0$ ($i \neq j$), therefore $\alpha_i(A) = \sum_{j=1}^n a_{ij}$. Further, for any $i = 1, \ldots, n$, considering the sign of $a_{ij}$ it is easy to calculate that

$$\alpha_i(A) = \sum_{j=1}^n a_{ij} = \sum_{j=1}^n \phi_i \phi_j = a(\phi_i, \sum_{j=1}^{n+n^\partial} \phi_j) - a(\phi_i, \sum_{j=n+1}^{n+n^\partial} \phi_j) =$$

$$a(\phi_i, 1) - a(\phi_i, \sum_{j=n+1}^{n+n^\partial} \phi_j) = \int_\Omega c\phi_i d\mathbf{x} - \sum_{j=n+1}^{n+n^\partial} a_{ij} \geq \int_\Omega c\phi_i d\mathbf{x} > 0,$$

where the very last strict inequality holds due to condition (2.12). Therefore, the matrix $A$ is always SDD for our type of problems. Moreover $A$ is an M-matrix, and therefore it is monotone [10]. Hence, the two-sided estimates (2.24) and (2.25) are valid.

The proofs of estimates (2.24) and (2.25) strongly depend on the computation of $a_{ij}$ and
Let us consider a simple case when \( c \) is constant and \( f \) is, e.g., piecewise polynomial so that all \( a_{ij} \) and \( F_j \) are computed exactly while implemented. We see immediately that, if all \( F_i \leq 0 \) then the first inequality in (2.24) and (2.25) holds. Let now some \( F_{i_0} > 0 \) for some index \( i_0 \in \{1, \ldots, n\} \) (i.e., \( \int_{\Omega} f \phi_{i_0} d\mathbf{x} > 0 \)). Then

\[
\frac{F_{i_0}}{\alpha_{i_0}(\mathbf{A})} \leq \frac{\int_{\Omega} f \phi_{i_0} d\mathbf{x}}{\int_{\Omega} c \phi_{i_0} d\mathbf{x}} \leq \frac{\int_{\Omega} \max\{0, \max_{\xi \in \Omega} f(\xi)\} \phi_{i_0} d\mathbf{x}}{c \int_{\Omega} \phi_{i_0} d\mathbf{x}} = \max\left\{0, \frac{\max_{\xi \in \Omega} f(\xi)}{c}\right\},
\]

and, similarly, if all \( F_{i_0} \geq 0 \), then the second inequality in (2.24) and (2.25) holds. Let now some \( F_{i_0} < 0 \) for some index \( i_0 \in \{1, \ldots, n\} \) (i.e., \( \int_{\Omega} f \phi_{i_0} d\mathbf{x} < 0 \)). Then

\[
\frac{F_{i_0}}{\alpha_{i_0}(\mathbf{A})} \geq \frac{\int_{\Omega} f \phi_{i_0} d\mathbf{x}}{\int_{\Omega} c \phi_{i_0} d\mathbf{x}} \geq \frac{\int_{\Omega} \min\{0, \min_{\xi \in \Omega} f(\xi)\} \phi_{i_0} d\mathbf{x}}{c \int_{\Omega} \phi_{i_0} d\mathbf{x}} = \min\left\{0, \frac{\min_{\xi \in \Omega} f(\xi)}{c}\right\},
\]

i.e., the estimates from (2.24) and (2.25) hold true in this case.

If \( c \) is not constant and \( f \) is not necessarily piecewise polynomial, then for the computations of the integrals \( \int_{\Omega} c \phi_i \phi_j d\mathbf{x} \) and \( \int_{\Omega} f \phi_j d\mathbf{x} \) in practice, we should use certain quadrature rules, and thus, each such case requires a separate analysis. In this work, we only demonstrate how to prove the required estimates if the simplest quadrature rule

\[
\int_S g(x) d\mathbf{x} \approx \frac{\text{meas}_d S}{N_S} \sum_{i=1}^{N_S} g(\xi_i) \tag{2.34}
\]

is used, where \( S \) is a finite element from the given mesh \( T_h \) and \( \xi_1, \ldots, \xi_{N_S} \) are its \( N_S \) vertices, furthermore, \( \text{meas}_d \) denotes a \( d \) dimensional measure. Then by using this formula we observe that

\[
a_{ij} = \int_{\Omega} \nabla \phi_i \cdot \nabla \phi_j d\mathbf{x} + \int_{\Omega} c \phi_i \phi_j d\mathbf{x} \approx \int_{\Omega} \nabla \phi_i \cdot \nabla \phi_j d\mathbf{x} \tag{2.35}
\]

for \( i \neq j \), and

\[
a_{ii} \approx \int_{\Omega} \nabla \phi_i \cdot \nabla \phi_i d\mathbf{x} + \frac{c(x_i) \text{meas}_d(\text{supp} \phi_i)}{N_S} > 0, \tag{2.36}
\]

where \( \text{supp} \phi_i \) denotes the support of the function \( \phi_i \). The sign \( \approx \) means that the value on the left-hand side of it is replaced in actual calculations by the value on the right-hand side of it, for the computing of which we use quadrature (2.34) and (2.28)–(2.30).

Therefore, if the finite element meshes and the basis functions are such that \( \nabla \phi_i \cdot \nabla \phi_j \leq 0 \) (which is less strict than the condition \( \nabla \phi_i \cdot \nabla \phi_j \leq -\varepsilon < 0 \) actually, required in the previous case [14], [17], [57]), we calculate that, if quadrature (2.34) is used for the practical
computing of the entries of $A$, then the actual values of numbers $\alpha_i(A)$ are such that

$$\alpha_i(A) \approx \frac{c(x_i) \text{ meas}_d(\text{supp } \phi_i)}{N_S}. \quad (2.37)$$

Further, using (2.34) again, now for the actual computing of the right-hand side of system (2.19), we see that

$$F_i \approx \frac{f(x_i) \text{ meas}_d(\text{supp } \phi_i)}{N_S}, \quad (2.38)$$

and it is now easy to show that the estimates in (2.24) and (2.25) do hold in this case, too.

**Remark 2.4.9** Some complicated finite element schemes, e.g., those leading to some positive off-diagonal entries but still to monotone matrices (see [10], [16], [63] or [79] for some examples) can be analyzed in the above manner. It is worth to mention here one interesting case, not covered by Windisch’s results [106], but provable due to Theorem 2.4.1 and analysis of the modified discrete maximum principle’s validity as in the previous examples of this section. Imagine that in some part of the solution domain $\Omega_0 \subset \Omega$ the righthand-side function $f$ is zero. Then it is natural that all entries $F_j$ associated with nodes $x_j$ lying in $\Omega_0$ are zeros in any meaningful schemes. Therefore, in estimates (2.22), the corresponding fractions $\frac{F_j}{\alpha_j(A)}$ are zeros a priori, independently of the values $\alpha_j(A)$. This means that the appearance of certain positive off-diagonal entries (with indices associate to $\Omega_0$) can be easily allowed without any effect on the desired two-sided estimations provided the resulting matrix $A$ remains SDD and monotone.

### 2.4.2 Applications for Finite Difference Methods

If, for example, the standard $(2d+1)$-node stencils [40] with uniform mesh of the step size $h$ in the available space directions are used, we get system (2.19) with a matrix which has strictly positive diagonal and nonpositive off-diagonal entries and which is always SDD, i.e., monotone [10]. As an illustration, such a type of schemes for the one-dimensional case and uniform mesh leads to the tridiagonal matrix

$$A = \text{tridiag} \left( -\frac{1}{h^2}, \frac{2}{h^2}, \frac{2}{h^2} + c(x_i), -\frac{1}{h^2} \right). \quad (2.39)$$

For such FD schemes (also in any dimensions), we always have $\alpha_i(A) \geq c(x_i)$ and $F_i = f(x_i)$, therefore, estimates (2.24) and (2.25) can be proved very easily again. Similarly to the finite element case we see immediately that if all $F_i \leq 0$, then the first inequality in (2.24) and (2.25) holds. Let now some $F_{i_0} > 0$ for some index $i_0 \in \{1, \ldots, n\}$. 20
Then
\[
\frac{F_{i_0}}{\alpha_{i_0}(A)} \leq \frac{f(x_{i_0})}{c(x_{i_0})} \leq \max \left\{ 0, \max_{\xi \in \Omega} f(\xi) \right\},
\]
and it is clear that if all \( F_{i_0} \geq 0 \), then the second inequality in (2.24) and (2.25) holds.

Let now some \( F_{i_0} < 0 \) for some index \( i_0 \in \{1, \ldots, n\} \). Then
\[
\frac{F_{i_0}}{\alpha_{i_0}(A)} \geq \frac{f(x_{i_0})}{c(x_{i_0})} \geq \min \left\{ 0, \min_{\xi \in \Omega} f(\xi) \right\},
\]
i.e., the estimates from (2.24) and (2.25) hold true in this case as well.

### 2.5 Maximum Principles for Problems with Third Boundary Conditions

In this section we consider the following boundary-value problem of elliptic type. Find a function \( u \in C^2(\Omega) \) such that
\[
-\Delta u + cu = f \quad \text{in } \Omega \quad \text{and} \quad \delta u + \frac{\partial u}{\partial \vec{n}} = g \quad \text{on } \partial \Omega,
\]
where \( \Omega \) is a bounded domain with Lipschitz continuous boundary \( \partial \Omega \), \( \vec{n} \) is the unit outward normal to \( \partial \Omega \), the reactive coefficient \( c(x) \geq 0 \) for all \( x \in \Omega \), and the coefficient \( \delta(s) \geq 0 \) for all \( s \in \partial \Omega \). The boundary condition in (2.40) is often called the third type boundary condition but also known as Newton or Robin boundary condition, see, e.g., [46].

In the followings we present and discuss the maximum principles for linear elliptic problems of second order with third boundary condition based on our results of the previous case with homogeneous Dirichlet condition. According to this, our key result for the maximum principle of this problem is the following [38].

**Theorem 2.5.1** Assume that in (2.40) the functions \( c, f \in C(\Omega) \), and the functions \( \delta, g \in C(\partial \Omega) \). In addition, let
\[
c(x) \geq c_0 > 0 \quad \text{for all } x \in \Omega \quad \text{and} \quad \delta(x) \geq \delta_0 > 0 \quad \text{for all } x \in \partial \Omega,
\]
where \( c_0 \) and \( \delta_0 \) are (positive) constants. Then the following (a priori) two-sided estimates for the classical solution of problem (2.40) are valid for any \( x \in \Omega \):
\[
\min \left\{ 0, \min_{\xi \in \Omega} \frac{f(\xi)}{c(\xi)}, \min_{\xi \in \partial \Omega} \frac{g(\xi)}{\delta(\xi)} \right\} \leq u(x) \leq \max \left\{ 0, \max_{\xi \in \Omega} \frac{f(\xi)}{c(\xi)}, \max_{\xi \in \partial \Omega} \frac{g(\xi)}{\delta(\xi)} \right\}.
\]
Proof: The first part of the proof is the same as in the Dirichlet case (see in Section 2.3), hence, only one the case, when \( u \) attains its positive maximum at some boundary point \( s_0 \in \partial \Omega \) gives us more work. Then \( \frac{\partial u}{\partial n}(s_0) \geq 0 \), and therefore from the boundary condition in (2.40) and the condition on \( \delta \) in (2.41) we obtain that \( u(s_0) \leq \frac{g(s_0)}{\delta(s_0)} \). From these considerations the upper estimate in (2.42) follows immediately. The lower estimate in (2.42) can be proved in the same way. ■

Similarly to the two-sided estimates (2.24) and (2.25) it is natural to give the following definition for this problem, too.

Definition We say that the solution \( u \) of system (2.19) with SDD matrix \( A \) satisfies the discrete maximum principle corresponding to the continuous maximum principle (2.42) if estimates (2.22) are valid and if, in addition, the estimates

\[
\max_{j=1,\ldots,N} \frac{F_j}{\alpha_j(A)} \leq \max \left\{ 0, \max_{x \in \Omega} \frac{f(x)}{c(x)}, \max_{s \in \partial \Omega} \frac{g(s)}{\delta(s)} \right\},
\]

(2.43)

\[
\min_{j=1,\ldots,N} \frac{F_j}{\alpha_j(A)} \geq \min \left\{ 0, \min_{x \in \Omega} \frac{f(x)}{c(x)}, \min_{s \in \partial \Omega} \frac{g(s)}{\delta(s)} \right\},
\]

(2.44)

are valid, i.e., the numerical solution remains between the same bounds as the continuous solution, or in other words the discretization is not increasing the continuous bounds.

Remark 2.5.1 In the case of earlier versions of continuous and discrete maximum principles no estimates like (2.24)–(2.25) were, in fact, needed as one dealt there with various implications involving the sign conditions only.

2.5.1 Applications for Finite Element Methods

Similarly to the homogeneous Dirichlet boundary case, in the sequel, we present how our theoretical results can be used for proving the modified discrete maximum principles for finite element method. Now the variational formulation of (2.40) for the finite element scheme, reads as follows. Find \( u \in H^1(\Omega) \) such that

\[
a(u, v) = F(v) \text{ for all } v \in H^1(\Omega),
\]

(2.45)

where

\[
a(u, v) = \int_\Omega \nabla u \cdot \nabla v \, dx + \int_\Omega cuv \, dx + \int_{\partial \Omega} \delta uv \, ds \quad F(v) = \int_\Omega fv \, dx + \int_{\partial \Omega} gv \, ds.
\]

(2.46)
The existence and uniqueness of the weak solution $u$ is provided by the Lax-Milgram lemma, the Friedrichs-type inequalities, and assumptions (2.41), see, e.g., [69]. It is worth emphasizing that in this case, actually, for the well-posedness of problem (2.45), one can require only that $c \in L^\infty(\Omega)$, $f \in L^2(\Omega)$, $\delta \in L^\infty(\partial \Omega)$, $g \in L^2(\partial \Omega)$.

Let $T_h$ be the same finite element mesh as in Section 2.4.1. Further, let the basis functions $\phi_1, \phi_2, \ldots, \phi_{n+n^0}$, associated with the given nodes have the same properties also (see (2.28)–(2.30)). These basis functions $\phi_1, \phi_2, \ldots, \phi_{n+n^0}$ are spanning a finite-dimensional subspace $V_h$ of $H^1(\Omega)$.

Now the finite element approximation of $u$ is defined as a function $u_h \in V_h$ such that

$$a(u_h, v_h) = \mathcal{F}(v_h) \quad \text{for all } v_h \in V_h,$$

whose existence and uniqueness are also provided by the Lax-Milgram lemma.

**Remark 2.5.2** Algorithmically, $u_h = \sum_{i=1}^{n+n^0} u_i \phi_i$, where the coefficients $u_i$ are the entries of the solution $u$ of system (2.19) with $a_{ij} = a(\phi_i, \phi_j)$, $F_i = \mathcal{F}(\phi_i)$, and $N = n + n^0$. It is clear that, if properties (2.28)–(2.30) hold, the finite element approximation $u_h$ satisfies the bounds from (2.22) at each point of $\overline{\Omega}$ if all its nodal values $u_i$ do satisfy them.

**Lemma 2.5.1** Assume that, based on the case with homogeneous Dirichlet boundary condition, the problem (2.40), under condition (2.41) on the coefficients is solved by finite element method with basis functions having properties (2.28)–(2.30). In addition, let matrix $A$ in the resulting matrix equation $Au = F$ be such that $a_{ij} \leq 0$ ($i \neq j$). Then $A$ is SDD and the estimates (2.22) are valid.

**Proof:** Clearly, from (2.46) and (2.41) it follows that $a_{ii} = a(\phi_i, \phi_i) > 0$ for all $i = 1, \ldots, n + n^0$. If $a_{ij} \leq 0$ ($i \neq j$), we observe for $i = 1, \ldots, n + n^0$ that

$$\alpha_i(A) = \sum_{j=1}^{n+n^0} a_{ij} = a(\phi_i, \sum_{j=1}^{n+n^0} \phi_j) = a(\phi_i, 1) = \int_\Omega c\phi_i d\mathbf{x} + \int_{\partial\Omega} \delta \phi_i d\mathbf{s} > 0,$$

where the last (strict) inequality holds due to (2.41). (We notice that, in fact, $\alpha_i(A) = \int_\Omega c\phi_i d\mathbf{x}$ if $i \in \{1, \ldots, n\}$.) Thus, the matrix $A$ is always SDD for our type of problems. Moreover $A$ is an M-matrix, and therefore it is monotone [10]. Hence, estimates (2.22) are valid, due to Theorem 2.4.1 with $\alpha_i(A)$ computed as in (2.48).

As we have seen previously, the proofs of estimates (2.43) and (2.44) strongly depend on how we compute $a_{ij}$ and $F_j$ in real finite element calculations. We consider in detail the following representative case.
Theorem 2.5.2 Assume that the coefficients $c$ and $\delta$ are (positive) constants and the functions $f$ and $g$ are such (e.g., piecewise polynomials) that all entries $a_{ij}$ and $F_j$ in system (2.19) are computed exactly. Then estimates (2.43) and (2.44) are valid provided by $a_{ij} \leq 0$ ($i \neq j$).

Proof: We see immediately that if $F_i \leq 0$ for all $i = 1, \ldots , n + n^\alpha$, then the upper estimate in (2.43) holds. Let now $F_{i_0} > 0$ for some index $i_0 \in \{1, \ldots , n\}$. Then due to (2.46) and (2.48) we observe

$$\frac{F_{i_0}}{\alpha_{i_0}(A)} = \frac{\int_{\Omega} f_{i_0} d\mathbf{x}}{\int_{\Omega} c_{i_0} d\mathbf{x}} \leq \frac{\int_{\Omega} \max\{0, \max_{\xi \in \Omega} f(\xi)\} \phi_{i_0} d\mathbf{x}}{c \int_{\Omega} \phi_{i_0} d\mathbf{x}} = \max\left\{0, \max_{\xi \in \Omega} \frac{f(\xi)}{c}\right\}.$$ 

Let now $F_{i_0} > 0$ for some index $i_0 \in \{n + 1, \ldots , n + n^\alpha\}$. Then, in view of (2.48), (1.23), and (2.41), we get

$$\frac{F_{i_0}}{\alpha_{i_0}(A)} = \frac{\int_{\Omega} f_{i_0} d\mathbf{x}}{\int_{\Omega} c_{i_0} d\mathbf{x}} \leq \frac{\int_{\Omega} \max\{0, \max_{\xi \in \Omega} f(\xi)\} \phi_{i_0} d\mathbf{x} + \int_{\partial \Omega} g_{i_0} d\mathbf{s}}{c \int_{\Omega} \phi_{i_0} d\mathbf{x} + \int_{\partial \Omega} \delta \phi_{i_0} d\mathbf{s}} \leq \max\left\{0, \max_{\xi \in \Omega} \frac{f(\xi)}{c}, \max_{\xi \in \partial \Omega} \frac{g(\xi)}{\delta}\right\}. \quad (2.49)$$

Similarly, if $F_i \geq 0$ for all $i = 1, \ldots , n + n^\alpha$, then the lower estimate in (2.44) holds. Let now $F_{i_0} < 0$ for some index $i_0 \in \{1, \ldots , n\}$. Then,

$$\frac{F_{i_0}}{\alpha_{i_0}(A)} = \frac{\int_{\Omega} f_{i_0} d\mathbf{x}}{\int_{\Omega} c_{i_0} d\mathbf{x}} \geq \frac{\int_{\Omega} \min\{0, \min_{\xi \in \Omega} f(\xi)\} \phi_{i_0} d\mathbf{x}}{c \int_{\Omega} \phi_{i_0} d\mathbf{x}} = \min\left\{0, \min_{\xi \in \Omega} \frac{f(\xi)}{c}\right\}. \quad (2.50)$$

Let now $F_{i_0} < 0$ for some index $i_0 \in \{n + 1, \ldots , n + n^\alpha\}$. Then, in view of (2.48), (1.23), and (2.41), we observe that

$$\frac{F_{i_0}}{\alpha_{i_0}(A)} = \frac{\int_{\Omega} f_{i_0} d\mathbf{x}}{\int_{\Omega} c_{i_0} d\mathbf{x}} \geq \frac{\int_{\Omega} \min\{0, \min_{\xi \in \Omega} f(\xi)\} \phi_{i_0} d\mathbf{x} + \int_{\partial \Omega} g_{i_0} d\mathbf{s}}{c \int_{\Omega} \phi_{i_0} d\mathbf{x} + \int_{\partial \Omega} \delta \phi_{i_0} d\mathbf{s}} \geq \min\left\{0, \min_{\xi \in \Omega} \frac{f(\xi)}{c}, \min_{\xi \in \partial \Omega} \frac{g(\xi)}{\delta}\right\}. \quad (2.51)$$

$$\frac{F_{i_0}}{\alpha_{i_0}(A)} = \frac{\int_{\Omega} f_{i_0} d\mathbf{x}}{\int_{\Omega} c_{i_0} d\mathbf{x}} \geq \frac{\int_{\Omega} \min\{0, \min_{\xi \in \Omega} f(\xi)\} \phi_{i_0} d\mathbf{x} + \int_{\partial \Omega} g_{i_0} d\mathbf{s}}{c \int_{\Omega} \phi_{i_0} d\mathbf{x} + \int_{\partial \Omega} \delta \phi_{i_0} d\mathbf{s}} \geq \min\left\{0, \min_{\xi \in \Omega} \frac{f(\xi)}{c}, \min_{\xi \in \partial \Omega} \frac{g(\xi)}{\delta}\right\}. \quad (2.52)$$

$\blacksquare$
Remark 2.5.3 If $c$ is not constant and $f$ is not necessarily piecewise polynomial, then for the computations of the entries (which are sums of integrals over $\Omega$ and its boundary $\partial \Omega$) in the system (2.19), we should, in practice, use certain quadrature rules, and, thus, each of such cases may require a separate analysis. For example, we proved the validity of the estimates (2.43) and (2.44) in our case if a simple quadrature rule considered in [36] is used in Section 2.4.1.

2.5.2 Applications for Finite Difference Methods

In this chapter, based on several representative finite difference schemes, we will demonstrate how the discrete maximum principle can be proved for finite difference methods by using the definition (2.43) and (2.44).

First, we consider the problem with Robin boundary condition (2.40) imposed in the one-dimensional domain $\Omega = (0,1)$. For the governing equation we will apply the following standard finite difference discretization:

$$
-y_{i-1} + \frac{2y_i - y_{i+1}}{h^2} + c_i y_i = f_i,
$$

where $i = 1, \ldots, \hat{n} - 1$, and we are using the step size $h = 1/\hat{n}$, while $c_i$ and $f_i$ denote the values of functions $c$ and $f$, respectively, and $y_i$ denotes the value of the unknown function at the node $ih$. The associated boundary conditions are discretized as follows:

$$
\delta_0 y_0 + \frac{y_0 - y_1}{h} = g_0
$$

and

$$
\delta_\hat{n} y_{\hat{n}} + \frac{y_{\hat{n}} - y_{\hat{n}-1}}{h} = g_{\hat{n}},
$$

where the lower index $j$ again means the value of the corresponding function at the node $jh$.

In the two-dimensional case let $y_{ij}$ denote the solution of the corresponding equation on the domain $\Omega = (0,1) \times (0,1)$, then using $h = 1/\hat{n}$ for step size in both directions and the classical 5-point stencil, the following equation holds inside the domain for all $i, j = 1, \ldots, \hat{n} - 1$:

$$
-\frac{y_{i-1,j} - y_{i+1,j}}{h^2} + 4y_{ij} - \frac{y_{i,j-1} - y_{i,j+1}}{h^2} + c_{i,j} y_{ij} = f_{i,j},
$$

where $c_{i,j}$ and $f_{i,j}$ denote the values of functions $c$ and $f$ respectively at the node $(ih, jh)$. The matrix of the system above, without the boundary conditions, is made of $n^* := (\hat{n}-1)^2$. 

25
rows and columns. The first order accurate finite difference discretization of the third boundary conditions is the following:

\[
\delta_{i,0}y_{i,0} + \frac{y_{i,0} - y_{i,1}}{h} = g_{i,0} \text{ for all } i = 1, 2, ..., \hat{n} - 1, \tag{2.60}
\]

\[
\delta_{i,\hat{n}}y_{i,\hat{n}} + \frac{y_{i,\hat{n}} - y_{i,\hat{n}-1}}{h} = g_{i,\hat{n}h} \text{ for all } i = 1, 2, ..., \hat{n} - 1, \tag{2.61}
\]

\[
\delta_{0,j}y_{0,j} + \frac{y_{0,j} - y_{0,j}}{h} = g_{0,j} \text{ for all } j = 1, 2, ..., \hat{n} - 1, \tag{2.62}
\]

\[
\delta_{\hat{n},j}y_{\hat{n},j} + \frac{y_{\hat{n},j} - y_{\hat{n}-1,j}}{h} = g_{\hat{n}h,j} \text{ for all } j = 1, 2, ..., \hat{n} - 1, \tag{2.63}
\]

where \(\delta_{i,j}\) and \(g_{i,j}\) denote the values of \(\delta\) and \(g\) at the node \((ih, jh)\), respectively. It is clear that the matrix of the boundary conditions are made of \(n^0 := 4\hat{n} - 1\) rows and columns. To summarize, we have a system of linear equations with \(N = n^* + n^0\) unknowns and the same number of equations. For the sake of simplicity we do not treat the corner points of the corresponding domain.

**Theorem 2.5.3** The finite difference discretization (2.59)–(2.63) has the following properties.

- It approximates a sufficiently smooth solution \(u\) with the first order of accuracy.
- The resulting finite difference matrix \(A\) is SDD and monotone.
- The estimates (2.43) and (2.44) are valid.

**Proof:** The first statement is obvious, hence, the approximation of the boundary condition (2.60)–(2.63) has first order accuracy. Moreover, it is easy to get the following. (For the sake of simplicity we will use single numeration for all the nodes in this proof as it should not lead to any misunderstanding, correspondingly we use notations like \(f_i\), \(c_i\), \(\delta_i\), etc. to denote the values of \(f\), \(c\), \(\delta\) etc. at a node with a (single) index \(i\)).

\[
\alpha_i(A) = c_i > 0, \text{ for all } i = 1, ..., n^* \tag{2.64}
\]

and

\[
\alpha_i(A) = \delta_i > 0, \text{ for all } i = n^* + 1, ..., n^0. \tag{2.65}
\]

Since, we assumed that \(\delta(x)\) and \(c(x)\) are positive functions, the matrix \(A\) is SDD and monotone, hence the second statement is valid. Further, for the right-hand side of the system we observe that

\[
F_i = f_i, \text{ for all } i = 1, ..., n^* \tag{2.66}
\]
and
\[ F_i = g_i, \quad \text{for all } i = n^* + 1, \ldots, n^0. \]  
(2.67)

Then for all \( i = 1, \ldots, n^* \) the following holds:
\[ \frac{F_i}{\alpha_i(A)} = \frac{f_i}{c_i} \leq \max \left\{ 0, \max_{i=1,\ldots,n^*} \frac{f_i}{c_i} \right\} \leq \max \left\{ 0, \max_{x \in \Omega} \frac{f(x)}{c(x)} \right\}. \]  
(2.68)

and for all \( i = n^* + 1, \ldots, n^0 \) we have
\[ \frac{F_i}{\alpha_i(A)} = \frac{g_i}{\delta_i} \leq \max \left\{ 0, \max_{i=n^*+1,\ldots,n^0} \frac{g_i}{\delta_i} \right\} \leq \max \left\{ 0, \max_{s \in \partial \Omega} \frac{g(s)}{\delta(s)} \right\}. \]  
(2.69)

Now it is easy to see that the relation
\[ \max_{i=1,\ldots,N} \frac{F_i}{\alpha_i(A)} \leq \max \left\{ 0, \max_{x \in \Omega} \frac{f(x)}{c(x)}, \max_{s \in \partial \Omega} \frac{g(s)}{\delta(s)} \right\}. \]  
(2.70)

holds, which immediately prove the upper estimate in (2.43). The calculations above for the lower estimate can easily be done on the base of (2.68) and (2.69).

\[ \text{Remark 2.5.4} \quad \text{Obviously, the same results as in the above theorem can be proved for the finite difference discretization (2.56)–(2.58).} \]

The previous approximation of the third boundary condition has only first order accuracy, which is not consistent with the second order accuracy of the finite difference discretization for the governing differential equation. Therefore, we will present and analyse another finite difference scheme, now with increased accuracy of the approximation for the third boundary condition. For simplicity, we discuss in detail only the more complicated two-dimensional case in this respect. The analysis of the one-dimensional case is similar, the exact higher order discretization of the third type boundary condition in one dimension is presented in the course of the numerical experiments (see Section 2.6). Let us approximate the third boundary condition in (2.40) on the boundary of the domain in the following way.

\[ \text{On the boundary } x = 0: \]
\[ \frac{y_{0,j} - y_{1,j}}{h} = \frac{h}{2} \left( \frac{y_{0,j+1} - 2y_{0,j} + y_{0,j-1}}{h^2} \right) + \frac{h}{2} \varepsilon_{0,j} y_{0,j} + \delta_{0,j} y_{0,j} = \]
\[ = g_{0,j} + \frac{h}{2} f_{0,j}, \quad j = 1, 2, \ldots, \hat{n} - 1, \]  
(2.71)
The finite difference discretization of (2.59), (2.71)–(2.74) has the following properties.

■ It approximates a sufficiently smooth solution \( u \) with the second order of accuracy.

■ The resulting finite difference matrix \( A \) is SDD and monotone.

■ The estimates (2.43) and (2.44) are valid.

Proof: We will show the first statement only for the boundary line \( x = 1 \). (The proofs of the other cases are similar.) Clearly, it is sufficient to show the second order of approximation at the boundary nodes only. Let us define

\[
\Psi_j = \frac{u(1, jh) - u(1 - h, jh)}{h} \quad - \frac{h}{2} \left( \frac{u(1, (j + 1)h) - 2u(1, jh) + u(1, (j - 1)h)}{h^2} \right) + \frac{h}{2} \frac{c_{i,j} u_{i,j} + \delta_{i,j} u_{i,j}}{2} - g(1, jh) - \frac{h}{2} f(1, jh),
\]

\( i = 1, 2, \ldots, n - 1, \)

\( j = 1, 2, \ldots, n - 1, \)

i.e., the local approximation error. Using the Taylor expansion, we get

\[
\frac{u(1, jh) - u(1 - h, jh)}{h} = \left( \partial_1 u - \frac{h}{2} \partial^2_1 u \right)_{(1, jh)} + \mathcal{O}(h^2),
\]

\( i = 1, 2, \ldots, n - 1, \)

\( j = 1, 2, \ldots, n - 1, \)
\[
\frac{u(1, (j + 1)h) - 2u(1, jh) + u(1, (j - 1)h)}{h^2} = (\partial_2^2 u)_{(1,jh)} + O(h^2),
\]
(2.77)

where the symbols like \( \partial_1 u \) and \( \partial_2 u \) denote the partial derivatives of \( u \) as usual. Hence, putting (2.76) and (2.77) into (2.75), we obtain

\[
\Psi_j = (\partial_1 u + \delta u - g)_{(1,jh)} - \frac{h}{2} \left( \partial_1^2 u + \partial_2^2 u + cu - f \right)_{(1,jh)} + O(h^2).
\]
(2.78)

Since \( \frac{\partial u}{\partial n}(1, y) = \partial_1 u(1, y) \), therefore, due to the boundary condition in (2.40), the first term on the right side vanishes. The second term also is equal to zero, because we have assumed that the solution satisfies the differential equation on the boundary, too. This shows the second order approximation.

For the second statement it is obvious that (2.64) holds, therefore to prove the statement, it is enough to show the diagonal dominance at the boundary nodes only. At these nodes we have (with single indices):

\[
\alpha_i(A) = \frac{h}{2} c_i + \delta_i > 0, \quad \text{for all } i = n^* + 1, \ldots, n^* + n^0.
\]
(2.79)

Therefore, under our assumptions \( A \) is SSD and, due to its sign-structure, it is also an M-matrix, and therefore monotone.

To prove the last statement, one could observe that for the right-hand side of the resulting finite difference system we have

\[
F_i = f_i \text{ for all } i = 1, \ldots, n^*, \quad \text{and } F_i = g_i + \frac{h}{2} f_i \text{ for all } i = n^* + 1, \ldots, n^* + n^0.
\]
(2.80)

Due to the second property, Theorem 2.5.1 can be used. Obviously, at the interior nodes the estimation will remain the same (2.69), however, at the boundary nodes (i.e., for \( i = n^* + 1, \ldots, n^* + n^0 \)) we observe, due to (1.23), that

\[
\frac{F_i}{\alpha_i(A)} = \frac{g_i + f_i h/2}{\delta_i + c_i h/2} \leq \max \left\{ 0, \max_{k=n^*+1,\ldots,n^*+n^0} \frac{g_k}{\delta_k}, \max_{k=1,\ldots,n^*+n^0} \frac{f_k}{c_k} \right\} 
\leq \max \left\{ 0, \max_{s \in \partial \Omega} \frac{g(s)}{\delta(s)}, \max_{x \in \Omega} \frac{f(x)}{c(x)} \right\}.
\]
(2.81)

The inequalities (2.81) and (2.69) prove the upper estimate (2.43). Similarly, one can prove the lower estimate (2.44) as well. 

\[\Box\]
2.6 Numerical Experiments

In this section we present and discuss the results of our numerical experiments conducted with MATLAB® for the one- and the two-dimensional problem with Robin boundary condition, for which the discretization schemes from the previous sections are used. The detailed numerical experiments for the homogeneous Dirichlet case can be found in [37].

The one-dimensional test problem reads as follows:

\[-\frac{d^2 u}{dx^2} + u = 4xe^x, \quad x \in \Omega = (0, 1),\]

\[\frac{\partial u}{\partial \vec{n}} + u(s) = g, \quad s \in \partial \Omega = \{0, 1\},\]

where \(g\) is a function defined as \(g(0) = -1\) and \(g(1) = -e\). The exact solution of this problem is \(u(x) = x(1 - x)e^x\).

For the discretization of this problem we use a linear finite element scheme and finite difference schemes discussed in the previous section. To this aim we divide the interval \(\Omega = (0, 1)\) into \(\hat{n}\) parts of equal length \(h\), i.e., \(h = 1/\hat{n}\).

For the second order approximation of boundary conditions (2.83) we use the following relations:

\[u_0 \left( \frac{1}{h} + \delta(0) + \frac{h}{2}c(0) \right) - u_1 \frac{1}{h} = g(0) + \frac{h}{2}f(0),\]

\[u_\hat{n} \left( \frac{1}{h} + \delta(1) + \frac{h}{2}c(1) \right) - u_{\hat{n}-1} \frac{1}{h} = g(1) + \frac{h}{2}f(1),\]

which, in our case, reduces to

\[u_0 \left( \frac{1}{h} + 1 + \frac{h}{2} \right) - u_1 \frac{1}{h} = -1,\]

\[u_\hat{n} \left( \frac{1}{h} + 1 + \frac{h}{2} \right) - u_{\hat{n}-1} \frac{1}{h} = e \left( 2h - 1 \right).\]

According to the exact solution of our problem we have the following bounds:

\[\min_{x \in \mathbb{H}} u(x) = 0, \quad \max_{x \in \mathbb{H}} u(x) = \left( \sqrt{5} - 2 \right) \exp \left( \frac{\sqrt{5} - 1}{2} \right) \approx 0.4380.\]
Table 2.1: Two-sided estimations of the first and second order finite difference methods

<table>
<thead>
<tr>
<th>$\hat{n}$</th>
<th>Estimated discrete MPR1</th>
<th>MPR2</th>
<th>Real discrete MPR1</th>
<th>MPR2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>min</td>
<td>max</td>
<td>min</td>
<td>max</td>
</tr>
<tr>
<td>10</td>
<td>-2.7183</td>
<td>8.6486</td>
<td>-2.0029</td>
<td>8.6486</td>
</tr>
<tr>
<td>100</td>
<td>-2.7183</td>
<td>10.4409</td>
<td>-2.6500</td>
<td>10.6551</td>
</tr>
<tr>
<td>1000</td>
<td>-2.7183</td>
<td>10.8297</td>
<td>-2.7115</td>
<td>10.8514</td>
</tr>
</tbody>
</table>

Table 2.2: Two-sided estimations of the linear finite element method

Calculations for the two-sided estimation in (2.42) lead to the bounds

$$u(x) \geq \min \left\{ 0, \min_{\xi \in \Omega} \frac{f(\xi)}{c(\xi)} \min_{\xi \in \partial \Omega} \frac{g(\xi)}{\delta(\xi)} \right\} = -e \approx -2.71828,$$

and

$$u(x) \leq \max \left\{ 0, \max_{\xi \in \Omega} \frac{f(\xi)}{c(\xi)} \max_{\xi \in \partial \Omega} \frac{g(\xi)}{\delta(\xi)} \right\} = 4e \approx 10.8731,$$

which is in agreement with (2.88).

Table 2.3 contains results of the two-sided estimations of the numerical solutions for the test problem (2.82)–(2.83) in the cases of the first (MPR1) and the second order (MPR2) finite difference schemes. The “real discrete” column presents the exact maximal and minimal values of the numerical solutions for different values of $\hat{n}$, while the “estimated discrete” column contains corresponding a priori bounds computed according to (2.22).

Table 2.3 illustrates the behavior of the error in the maximum norm for all the three numerical schemes used (MPR1, MPR2, MPR3), which supports the theoretical analysis on the rate of the convergence of different numerical methods.

<table>
<thead>
<tr>
<th>$\hat{n}$</th>
<th>MPR1</th>
<th>MPR2</th>
<th>MPR3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2.95E-01</td>
<td>1.61E-02</td>
<td>7.41E-02</td>
</tr>
<tr>
<td>100</td>
<td>2.74E-02</td>
<td>1.33E-04</td>
<td>8.88E-04</td>
</tr>
<tr>
<td>1000</td>
<td>2.72E-03</td>
<td>1.31E-06</td>
<td>9.04E-06</td>
</tr>
</tbody>
</table>

Table 2.3: Convergence of the applied numerical schemes
Table 2.4: Results of the first order finite difference method applied to the two-dimensional problem (MPR4)

For the two-dimensional case let us consider the following problem on the unit square.

\[-\Delta u + u = f(x, y), \quad (x, y) \in \Omega = (0, 1) \times (0, 1),\]
\[\frac{\partial u}{\partial n} + u(s) = g, \quad s \in \partial \Omega,\]

whose exact solution is defined as \(u(x, y) = xy(1-x)(1-y)e^{x+y}\). (The functions \(f\) and \(g\) can be correspondingly found then.)

We can easily compute that

\[
\min_{\Omega} u(x, y) = 0, \quad \max_{\Omega} u(x, y) = \left(\sqrt{5} - 2\right)^2 \exp\left(\sqrt{5} - 1\right) \approx 0.1918.
\]

Using the two-sided estimation from (2.42), we get the estimates

\[
u(x) \geq \min \left\{ 0, \min_{\xi \in \Omega} \frac{f(\xi)}{c(\xi)}, \min_{\xi \in \partial \Omega} \frac{g(\xi)}{\delta(\xi)} \right\} = -\left(\sqrt{5} - 2\right) \exp\left(\frac{\sqrt{5} + 1}{2}\right) \approx -1.1905
\]

and

\[
u(x) \leq \max \left\{ 0, \max_{\xi \in \Omega} \frac{f(\xi)}{c(\xi)}, \max_{\xi \in \partial \Omega} \frac{g(\xi)}{\delta(\xi)} \right\} \approx 4.9577,
\]

which is in agreement with (2.93).

In order to compute the approximations, we have used the first (MPR4) and the second order (MPR5) finite difference methods again as described before. Table 2.4 and Table 2.5 reveal the results of the two-sided estimation for the corresponding numerical solutions. According to the numbers the theoretical results are validated in this case as well.

**Remark 2.6.1** It is worth emphasizing that, according to the numerical result, by decreasing the spatial step size in the case of first order finite difference methods (in one and two dimensions as well), the discrete lower estimation tends to the continuous lower estimation faster than the second order finite difference methods. However, for the upper
Table 2.5: Results of second order finite difference method for the two-dimensional problem (MPR5)

<table>
<thead>
<tr>
<th>$\hat{n} \times \hat{n}$</th>
<th>Estimated discrete</th>
<th>Real discrete</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>min</td>
<td>max</td>
</tr>
<tr>
<td>$10 \times 10$</td>
<td>-9.10E-02</td>
<td>4.9536</td>
</tr>
<tr>
<td>$20 \times 20$</td>
<td>-5.31E-02</td>
<td>4.5970</td>
</tr>
<tr>
<td>$40 \times 40$</td>
<td>-2.82E-02</td>
<td>4.9559</td>
</tr>
<tr>
<td>$80 \times 80$</td>
<td>-1.45E-02</td>
<td>4.9577</td>
</tr>
</tbody>
</table>

Figure 2.1: The solution of problem (2.91) by applying the first order finite difference method (MPR4)

*estimation there is no significant difference between the two finite difference methods.*

In two dimensions for $\hat{n} = 40$, the first (Fig. 2.1) and the second (Fig. 2.2) order finite difference results are shown respectively. It can be seen that in the higher order case the solution is much smoother and it is closer to the exact solution as well, which is non-negative on the given domain.
Figure 2.2: The solution of problem (2.91) by applying the second order finite difference method (MPR5)
Chapter 3

Non-negativity Preservation for Parabolic Problems

Parabolic partial differential equations are often applied to construct such mathematical models of non-stationary phenomena as the heat conduction or diffusion processes, reaction-diffusion problems (air pollution models, e.g., [108]), problems of electrodynamics (Maxwell equations, see, e.g., [97]), option pricing models (Black-Scholes models [9]), and many others in different fields of biology, chemistry, economy, sociology, etc.

In this chapter we consider the heat conduction equation as a typical prototype for parabolic type differential equations. It is clear that the temperature in a given domain cannot be negative if the temperature was non-negative initially and was kept non-negative on the boundary of the domain as well. This property is called the non-negativity preservation [30].

However, the non-negativity preservation is only a special case of the more general property called maximum principle (see the previous chapter for elliptic problems). In the case of the heat conduction equation the maximum principle states that the temperature must be bounded by the initial temperature, the strength of the heat source and the temperature at the spatial boundary.

The discrete maximum principle for parabolic problems was originally discussed almost 30 years ago, see, e.g., the works of Fujii [43] and Keller [61]. Based on the acuteness of the tetrahedral meshes Fujii [43] in 1973 obtained sufficient conditions of the discrete maximum principle for the finite element solution of some parabolic problems. The lumped mass method and some hyperbolic problems are considered in [11]. In 2005 Faragó et. al [31] derived necessary and sufficient conditions of the discrete maximum principle for finite element methods and sufficient conditions were given for hybrid meshes. Comprehensive surveys on discrete maximum principles can be found in papers [13], [citefarhor2], [citefarhor3]. The conditions of the discrete non-negativity preservation was discussed in [28], [49] for linear finite elements in one, two and three dimensions, and in [25] for a one-dimensional case with the combination of the finite difference and finite element methods.
For nonlinear problems the discrete non-negativity preservation was investigated in [102]. During the numerical solution of PDE’s, preservation of the basic qualitative properties of the original or physical solution is a very important requirement, assuming that they are inherent to the continuous mathematical model. Hence, the temperature – measured in kelvin – is non-negative by definition, and our expectation for positive numerical results is natural.

In what follows, first the non-negativity preservation for the semidiscrete solutions of (3.4) will be analyzed and, additionally, direct connection between the non-negativity preservation of the semidiscrete solutions (3.4) for the one- and two-dimensional cases will be established.

In Section 3.3.2 the exact condition for the non-negativity preservation in one dimension and the bounds for the linear finite element method will be given. In Section 3.3.4, conditions under which the bilinear finite element method is non-negativity preserving are formulated [94], [95], [96]. For the non-negativity of the difference schemes, see, e.g. [35].

3.1 Model Problem

We consider the following model problem:

\[
\frac{\partial u}{\partial t} - \Delta^d u = 0 \quad \text{in} \quad \Omega_T = (0, T) \times \Omega, \\
\frac{\partial u}{\partial t} = 0 \quad \text{on} \quad \Gamma_T = (0, T) \times \partial \Omega, \\
u|_{t=0} = u_0 \quad \text{on} \quad \Omega,
\]

where \(u\) stands for the temperature of the solution domain, \(t\) denotes time, and \(u_0\) is a given initial function defined in \(\Omega\). According to the physical phenomena, the first term on the left-hand side of this equation expresses the rate of the temperature change at a point in space over time, and the second term on the left-hand side indicates the spatial thermal conduction. The physical meaning of the problem above is the heat conduction on the given domain (for example a rod in one dimension) without any heat source, at the boundary (at the ends of the rod) the temperature is absolute zero, and the initial temperature distribution (along the rod) is defined by \(u_0\).

For this equation, the non-negativity preservation principle holds and reads as follows: for any non-negative initial function \(u_0\), the solution \(u\) has to be non-negative in \(\Omega_T\) as well, see, e.g., [89]. This expectation is valid since the temperature in kelvin is a non-negative quantity in physics.
The most common numerical approach for solving the system (3.1)–(3.3) is the combination of separate discretizations in space and time. For the spatial one, we can apply the finite element method or the finite difference method with a given equidistant mesh size $h$. As a result, one can get the following Cauchy problem for the semidiscrete solution $u_h$, which is a time dependent vector-valued function with entries assigned to the nodes of the given mesh:

$$\frac{du_h}{dt}(t) = \Delta_h u_h(t), \ t \in (0, T),$$  \hfill (3.4)\]

where the initial value $u_h(0)$ is given, and $\Delta_h$ denotes the corresponding discrete Laplace operator (represented by a square matrix, see, e.g., (3.10) and (3.11)).

### 3.2 Semidiscrete Solutions

In this section, we consider the non-negativity preservation for the Cauchy problem (3.4), where $\Delta_h$ arises from the uniform space discretization of the Laplace operator on rectangular mesh. The solution has the form

$$u_h(t) = \exp(t \Delta_h) u_0, \ t \in (0, T),$$ \hfill (3.5)\]

where $\exp(t \Delta_h)$ denotes the exponent of the matrix $t \Delta_h$ (see (1.2)). Therefore, the problem of non-negativity preservation in this case is equivalent to finding conditions on the discretization methods under which the matrix exponential $\exp(t \Delta_h)$ is non-negative.

The following useful lemma holds [8].

**Lemma 3.2.1** Let $A$ be an arbitrary square matrix with the entries $a_{ij}$. Then $\exp(tA)$ is non-negative for any $t \geq 0$ if and only if the condition

$$a_{ij} \geq 0 \quad \text{for all } i \neq j$$ \hfill (3.6)\]

holds.

**Proof:** By the definition of the matrix exponential, we have

$$\exp(tA) = I + tA + ... ,$$ \hfill (3.7)\]

where $I$ denotes the appropriate identity matrix. This series immediately shows the necessity of condition (3.6), since for small values of $t$, the latter terms could not influence
the sign. Let now \( s \) be a scalar such that \( A + sI \) is a non-negative matrix. Then, obviously, \( \exp(t(A + sI)) \) is non-negative if \( t \geq 0 \). Moreover, \( \exp(-tsI) \) is also non-negative (it is a diagonal matrix, with the entries \( e^{(-st)} \), see (1.4)), and the matrices \( t(A + sI) \) and \( -tsI \) commute. Therefore, due to the identity

\[
\exp(tA) = \exp(t(A + sI) - tsI) = \exp(t(A + sI)) \cdot \exp(-tsI),
\]

(3.8)
the sufficiency of condition (3.6) is also proven.

\[\Box\]

3.2.1 One-Dimensional Case

If problem (3.1)–(3.3) is considered in one-dimensional setting, the structure of the matrix \( \Delta_h \) is well-known for both the finite difference and finite element discretizations. Namely, using the standard notations for the so-called stiffness and mass matrices of a size defined by the spatial mesh, respectively:

\[
Q = \text{tridiag}(1, -2, 1), \quad M = \frac{1}{6}\text{tridiag}(1, 4, 1),
\]

(3.9)
then, in case of the finite difference method

\[
\Delta_h = \frac{1}{h^2} Q,
\]

(3.10)
and for the linear finite element method

\[
\Delta_h = \frac{1}{h^2} M^{-1} Q.
\]

(3.11)
The matrix \( \Delta_h \) from (3.10) obviously satisfies condition (3.6), at the same time, the matrix \( \Delta_h \) from (3.11) (which can be computed explicitly) is known to have its entries changing the sign chessboard-like [33], i.e., it does not satisfy condition (3.6).

Thus, using Lemma 3.2.1, the following result is obtained.

**Theorem 3.2.1** For the one-dimensional problem (3.4), the semidiscrete numerical solutions, obtained by the finite difference discretization, preserve the non-negativity property. However, this property is not preserved, in general, by the numerical solutions resulting from the linear finite element semidiscretization.

**Remark 3.2.1** In [11] there is a condition for the positivity of the time derivative of the semidiscrete solutions. Obviously, this condition can be regarded as a sufficient condition for the non-negativity preservation of the semidiscrete solutions.
Remark 3.2.2 Let us introduce the following notation

\[ T_1(p) = \text{tridiag}(1, p, 1), \quad (3.12) \]

where \( p \in \mathbb{R} \), i.e., \( Q = T_1(-2) \). Then, the one-dimensional semidiscretization of (3.1) in the form

\[ \frac{du_h(t)}{dt} = \frac{1}{h^2} T_1(p)u_h(t), \quad t \in (0, T), \quad (3.13) \]

is non-negativity preserving for any value of the parameter \( p \). Therefore, instead of (3.13), a more general equation can be considered:

\[ \frac{\partial u}{\partial t} = \Delta d u + ku \quad \text{in} \quad \Omega_T = (0, T) \times \Omega, \quad (3.14) \]

where \( k(x) \) is a non-negative function, and it can be proved that the finite difference semidiscretization for such an equation is non-negativity preserving, since the approximation of the new term affects only the diagonal entries of the matrix in (3.13).

Remark 3.2.3 However, the non-negativity preservation property on continuous level is known to hold only for some \( k \geq k_0 \) (see, e.g., [89]).

3.2.2 Two-Dimensional Case

In this section the discretization on the uniform mesh (of step size \( h \)) of problem (3.1)–(3.3) in the two-dimensional case is considered. Based on (3.12) let us introduce the notation

\[ T_2(p) = \text{tridiag}(I, T_1(p), I) \quad (3.15) \]

for a block tridiagonal matrix (from \( \mathbb{R}^{n^2 \times n^2} \)), where \( p \in \mathbb{R} \).

Obviously, if the the standard finite difference method is applied, then in the corresponding Cauchy-problem (3.1), the matrix \( \Delta_h \) has the form

\[ \Delta_h = \frac{1}{h^2} T_2(-4), \quad (3.16) \]

that is, all its off-diagonal entries are non-negative. By using the linear finite element method, we get \( \Delta_h = (1/h^2)M_2^{-1}T_2(-4) \), where \( M_2 \) denotes the corresponding mass matrix:

\[ M_2 = \frac{1}{12} \text{tridiag}(M_d^T, T_1(6), M_d), \quad (3.17) \]
where

\[ M_d = \text{tridiag}(0, 1, 1). \]  (3.18)

In what follows, the relation between the exponentials of the matrices \( T_1(p) \) from \( \mathbb{R}^{n \times n} \) and \( T_2(p) \) from \( \mathbb{R}^{n^2 \times n^2} \) is established, i.e., the matrix exponential between the 1D and 2D discrete Laplacians. Using the relation

\[ A(p) = \text{tridiag}(1, p + 2, 1) = T_1(p) + 2I, \]  (3.19)

we have, see, e.g., [39, 85],

\[ T_2(p) = \text{tridiag}(I, A(p) - 2I, I) = \text{tridiag}(0, A(p), 0) + \text{tridiag}(I, -2I, I) = I \otimes A(p) + Q \otimes I = Q \oplus A(p), \]  (3.20)

where \( \otimes \) and \( \oplus \) denotes the Kronecker product and the sum of matrices, respectively, see (1.8) and (1.9). In order to attribute the matrix exponential of the matrix \( T_2(p) \in \mathbb{R}^{n^2 \times n^2} \) to the matrix exponentials of the matrices \( A(p) \) and \( Q \) from \( \mathbb{R}^{n \times n} \), i.e., the two-dimensional problem to the one-dimensional, we prove the following lemma.

**Lemma 3.2.2** For the matrices \( T_2(p), A(p), \) and \( Q \), the relation

\[ \exp(T_2(p)) = \exp(Q) \otimes \exp(A(p)) \]  (3.21)

holds.

**Proof:** For any matrices \( A, B, C, D \) of the same size, we have [39, p. 228]

\[ (A \otimes B)(C \otimes D) = AC \otimes BD. \]  (3.22)

Therefore

\[ (I \otimes A(p))(Q \otimes I) = Q \otimes A(p), \]
\[ (Q \otimes I)(I \otimes A(p)) = Q \otimes A(p). \]  (3.23)

Consequently, the corresponding exponential can be written by use of the binomial rule as follows
\begin{align*}
\exp(T_2(p)) &= \exp((I \otimes A(p)) + (Q \otimes I)) = \\
&= \sum_{n=0}^{\infty} \frac{1}{n!} (I \otimes A(p) + Q \otimes I)^n = \\
&= \sum_{n=0}^{\infty} \sum_{k=0}^{n} \frac{1}{k!(n-k)!} (I \otimes A(p))^k (Q \otimes I)^{n-k} = \\
&= \sum_{n=0}^{\infty} \sum_{k=0}^{n} \frac{1}{k!(n-k)!} Q^k \otimes A(p)^{n-k}.
\end{align*}

On the other hand, by the definition of the tensor product, we have

\begin{align*}
\exp(Q) \otimes \exp(A(p)) = \sum_{i=0}^{\infty} \frac{1}{i!} Q^i \otimes \sum_{j=0}^{\infty} \frac{1}{j!} A(p)^j &= \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \frac{1}{i!j!} Q^j \otimes A(p)^j. 
\end{align*}

Since the right-hand sides in (3.24) and (3.25) are equal, we obtain the relation (3.21).

Obviously, the tensor product of two matrices is non-negative if and only if both involved matrices are non-negative. Moreover, \(\exp(Q)\) is non-negative and \(\exp(A(p))\) is non-negative if and only if \(\exp(T_1(p))\) is non-negative. Using Lemma 3.2.1, we obtained the following statement.

**Theorem 3.2.2** For the two-dimensional problem (3.1)–(3.3) defined on rectangular mesh, the semidiscrete numerical solution, obtained by the regular finite difference discretization, preserves the non-negativity property. However, this property is not preserved, in general, for the linear finite element discretization.

**Remark 3.2.4** The statement of Theorem 3.2.2 is also valid for the more general equation (3.14).

## 3.3 Qualitative Properties of the Fully Discretized Numerical Model

In this section the non-negativity preservation property of the fully discretized one-dimensional heat conduction problem (3.1)–(3.3) with homogeneous Dirichlet boundary conditions (3.2) is considered.
Applying the so-called $\theta$-method (or weighted method) with the given time step $\tau$ and the numerical parameter $\theta \in [0, 1]$ to problem (3.4), the following algebraic iterative equation presents the fully discretized problem

$$X_1 y^{\ell+1} = X_2 y^{\ell},$$  

(3.26)

where $X_1$ and $X_2$ are given $n$-by-$n$ matrices ($n$ denotes the ordinal number of nodes), and the vector $y^{\ell}$ represents the approximation to the vector $u_h(\ell\tau)$, moreover $y^0 = u_h(0)$ represents the given initial condition. It is clear that the matrices $X_1$, $X_2$ represent the applied numerical method, hence, their exact form will be given for our problem in latter sections that are dedicated to the non-negativity preservation property of the fully discretized problem solved by finite difference and finite element methods.

Our aim is to formulate, for a fixed parameter $\theta$, such conditions on the discretization parameters $h$ and $\tau$, under which the corresponding fully discretized problems preserve the non-negativity property.

In the next sections it will be shown that in the one-dimensional case, the exact (necessary and sufficient) conditions for the non-negativity preservation (with respect to $\tau$ and $h$, for any fixed $\theta$) can be obtained. The results are based on finding the exact representation of the matrix $X = X_1^{-1}X_2$, where the crucial point is to find the inverse matrix $X_1^{-1}$ if it exists.

In the two-dimensional case, a similar problem is more difficult, due to more complicated structures of the corresponding matrices (which are now block matrices). Some sufficient condition is given in [28] and it is based on the following requirements: $X_2 \geq 0$ and $X_1$ is a monotone matrix. However, finding the necessary and sufficient conditions is still an open problem.

After performing the full discretization of problem (3.4) we get equation (3.26) with matrices from $\mathbb{R}^{n \times n}$ in the form:

$$X_1 = \frac{1}{\tau}M - \theta Q, \quad X_2 = \frac{1}{\tau}M + (1 - \theta)Q,$$  

(3.27)

where $M = I$ for the finite difference method, and it has the form (3.9) for the linear finite element method. Therefore, for the finite difference method we have
\[ X_1 = \text{tridiag} \left( -\frac{\theta}{h^2}, \frac{1}{\tau} + 2\frac{\theta}{h^2}, \frac{\theta}{h^2} \right), \] (3.28)

\[ X_2 = \text{tridiag} \left( \frac{1-\theta}{h^2}, \frac{1}{\tau} - 2\frac{1-\theta}{h^2}, \frac{1-\theta}{h^2} \right), \] (3.29)

and for the linear finite element method \[96\] the corresponding matrices are

\[ X_1 = \text{tridiag} \left( \frac{1}{6\tau} - \frac{\theta}{h^2}, \frac{2}{3\tau} + \frac{2\theta}{h^2}, \frac{1}{6\tau} - \frac{\theta}{h^2} \right), \] (3.30)

\[ X_2 = \text{tridiag} \left( \frac{1}{6\tau} + \frac{1-\theta}{h^2}, \frac{2}{3\tau} - \frac{2(1-\theta)}{h^2}, \frac{1}{6\tau} + \frac{1-\theta}{h^2} \right). \] (3.31)

For the non-negativity preservation property we have to require the condition

\[ X = X_1^{-1}X_2 \geq 0, \] (3.32)

where \( X \) is the so-called iteration matrix.

Let us notice that the matrices in (3.28) and (3.30) have special structure: only the entries of the main-, super- and subdiagonals differ from zero, and the entries standing on the same diagonal are equal. Moreover, these matrices are symmetric. Such kind of matrix is called uniformly continuant symmetrical tridiagonal matrix, which has some special qualitative properties, considered in the sequel.

### 3.3.1 Non-negativity of the Iteration Matrix

The real, uniformly continuant symmetrical tridiagonal matrices are considered with \( z, w, s, p \in \mathbb{R} \)

\[ X_1 = z \cdot \text{tridiag}(-1, 2w, -1); \quad X_2 = s \cdot \text{tridiag}(1, p, 1) \] (3.33)

with the assumptions

\[ z > 0, \quad s > 0, \quad w > 1. \] (3.34)

The aim is to define those conditions under which the matrix \( X = X_1^{-1}X_2 \) is non-negative. Let us introduce the following so-called one-pair matrix \( G = (G_{ij}) \) (see (1.10)):
\[ G_{i,j} = \begin{cases} \gamma_{i,j}, & \text{if } i \leq j \\ \gamma_{j,i}, & \text{if } j \leq i \end{cases} \quad (3.35) \]

where \((i, j = 1, 2, \ldots, n)\) and
\[ \gamma_{i,j} = \frac{\text{sh}(iv)\text{sh}((n + 1 - j)v)}{\text{sh}(v)\text{sh}((n + 1)v)} \quad (3.36) \]

where \(v = \text{arch}(w)\) with \(w > 1\). We note that the matrix \(G\) is the function of the scalar \(w\).

Due to the relation \(X_1^{-1} = (1/z)G\) (see [85]), the following lemma is valid.

**Lemma 3.3.1** For the matrices \(X_1\) and \(X_2\) of the form \((3.33)\) the iteration matrix \(X = X_1^{-1}X_2\) can be expressed as
\[ X = \frac{s}{z}[(2w + p)G - I]. \quad (3.37) \]

**Proof:** According to the definitions of \(X_1\) and \(X_2\), the equation
\[ X_2 = s(2w + p)I - \frac{s}{z}X_1 \quad (3.38) \]
holds. By rearranging the terms and multiplying the equation \((3.38)\) by \(G = X_1^{-1}/z\) on the left side we get
\[ X := X_1^{-1}X_2 = \frac{s}{z}[(2w + p)G - I], \quad (3.39) \]
which proves the lemma.

Hence, taking into account the conditions \((3.34)\) and relation \((3.37)\), we get the following statement.

**Lemma 3.3.2** Under the condition \((3.34)\) the matrix \(X \in \mathbb{R}^{n \times n}\) is non-negative if and only if the conditions
\[ 2w + p > 0 \quad (3.40) \]
and
\[ \gamma_{i,i} \geq \frac{1}{2w + p}, \quad i = 1, 2, \ldots, n \quad (3.41) \]
are fulfilled.

Now let us analyze the expression on the left-hand side of condition \((3.41)\).
Lemma 3.3.3 For the entries in (3.36), with \( i = j \), the relation
\[
\min \{ \gamma_{i,i}, \ i = 1, 2, \ldots, n \} = \gamma_{1,1} = \gamma_{n,n} \tag{3.42}
\]
holds.

Proof: Introducing the functions \( h_1(y) = K_1 \sh(Cy)\sh(C(n+1-y)) \) and \( h_2(y) = K_2 y(n+1-y) \) on the interval \([1, n]\), (where \( K_1, K_2 \) and \( C \) are some positive real constants), considering the monotonicity of the hyperbolic sine, one can check that both functions \( h_1(y) \) and \( h_2(y) \) assume their maxima at the same point \( y = (n+1)/2 \). Moreover, on the interval \([1, (n+1)/2]\) they are monotonically increasing, while on the interval \(((n+1)/2, n]\) they are monotonically decreasing. Using this fact and the expressions for \( \gamma_{i,i} \), we get the statement. ■

Combining Lemma 3.3.2 and Lemma 3.3.3 we obtain

Theorem 3.3.1 Under conditions (5.34), for arbitrary fixed \( n \) the matrix \( X \in \mathbb{R}^{n \times n} \) is non-negative if and only if conditions (3.40) and
\[
a(n) := \frac{\sh(n\vartheta)}{\sh((n+1)\vartheta)} \geq \frac{1}{2w+p} \tag{3.43}
\]
are satisfied.

Obviously, (3.40) and (3.43) are necessary and sufficient conditions of the non-negativity for some fixed dimension \( n \). Let us turn to the examination of the varying \( n \). Due to the relations
\[
\frac{\sh(n\vartheta)}{\sh((n+1)\vartheta)} = \ch(\vartheta - \coth((n+1)\vartheta)\sh(\vartheta)), \tag{3.44}
\]
and the fact, that \( \coth(\alpha) \) tends to 1 when \( \alpha \) tends to infinity, we have
\[
\sup \left\{ \frac{\sh(n\vartheta)}{\sh((n+1)\vartheta)} : \ n \in \mathbb{N} \right\} = \ch(\vartheta) - \sh(\vartheta) = \exp(-\vartheta). \tag{3.45}
\]
Since the sequence \( a(n) \) is monotonically increasing, it converges to its limit (which is its superior) monotonically. Thus, conditions (3.40) and (3.43), i.e., the necessary and sufficient conditions for some fixed \( n \), serve as sufficient condition of the non-negativity of the matrices \( X \in \mathbb{R}^{n_1 \times n_1} \) for all \( n_1 \geq n \). Let us observe that
\[
\exp(-\vartheta) = \exp(-\arch(w)) = \exp \left( \ln \left[ w + \sqrt{w^2 - 1} \right]^{-1} \right) = \left[ w + \sqrt{w^2 - 1} \right]^{-1}. \tag{3.46}
\]
Therefore, for some sufficiently large $n^* \in \mathbb{N}$, the relation $X \geq 0$ may be true only if the condition

$$\left[w + \sqrt{w^2 - 1}\right]^{-1} > \frac{1}{2w + p}, \quad (3.47)$$

i.e., the condition

$$p > -w + \sqrt{w^2 - 1} \quad (3.48)$$

is fulfilled and $G > 0$. This leads to the following statement

**Theorem 3.3.2** Assume that the conditions in (3.34) are satisfied. If, for some number $n_0 \in \mathbb{N}$, conditions (3.40) and (3.43) are satisfied, then all matrices $X \in \mathbb{R}^{n \times n}$ with $n \geq n_0$ are also non-negative. Moreover, there exists such a number $n_0$ if and only if the condition (3.48) holds.

**Remark 3.3.1** Since

$$a(1) = \frac{\text{sh}\vartheta}{\text{sh}(2\vartheta)} = \frac{1}{2\text{ch}\vartheta} = \frac{1}{2w}, \quad (3.49)$$

therefore, (3.43) results in the condition

$$p \geq 0. \quad (3.50)$$

**Remark 3.3.2** Due to the relation

$$a(2) = \frac{\text{sh}(2\vartheta)}{\text{sh}(3\vartheta)} = \frac{2\text{ch}(\vartheta)}{4\text{ch}^2(\vartheta) - 1} = \frac{2w}{4w^2 - 1}, \quad (3.51)$$

condition (3.43) results in the assumption

$$p \geq -\frac{1}{2w}. \quad (3.52)$$

Therefore, $X \in \mathbb{R}^{n \times n}$ is non-negative for all $n = 2, 3, \ldots$ if and only if $X_1$ is an M-matrix, and according to the monotonicity of the sequence $a(n)$ (3.52) is valid.

**Remark 3.3.3** Conditions (3.50) and (3.52) (corresponding to the cases $n = 1$ and $n = 2$, respectively) are sufficient conditions for the non-negativity of the matrix $X$ for any larger size. For increasing $n$, the new conditions, which we obtain, are approaching the necessary condition of non-negativity. Using (3.43) and (3.44) we can characterize the rate of the convergence: it is equal to the rate of convergence of the sequence \{coth($n\vartheta$), $n = 1, 2, \ldots$\} to one. Based on the well-known definition of coth($x$),
\[ \coth(n \vartheta) = 1 + \frac{2}{[\exp(\vartheta)]^{2n}} - 1. \quad (3.53) \]

Using (3.46),

\[ \exp(\vartheta) = w + \sqrt{w^2 - 1} =: \beta. \quad (3.54) \]

Hence, the sequence of the bounds of the sufficient conditions converges linearly with the ratio \( 1/\beta^2 \) to the bound of the necessary condition.

### 3.3.2 Non-negativity of Finite Element Schemes

In this section we consider the non-negativity preservation property for the finite element discretization of the one and two-dimensional heat conduction equation. First, by analyzing the linear discretization of the one-dimensional equation (3.1) the corresponding matrices of the linear finite element method are (3.30) and (3.31). Based on the notation of the previous section, it is easy to obtain the constants

\[ w = \frac{1}{3\tau} + \frac{\theta}{h^2} + \frac{1}{1 - \frac{h^2}{6\tau}}, \quad p = -\frac{2}{1} \left( \frac{1 -\theta}{h^2} - \frac{1}{3\tau} \right). \quad (3.55) \]

To satisfy the conditions of Theorem 3.3.1 we need \( w > 1 \), which now can only be fulfilled if

\[ \frac{h^2}{6\theta} > \tau. \quad (3.56) \]

Moreover, according to Remark 3.3.2, for the non-negativity preservation we need \( p \geq 0 \), i.e.,

\[ \frac{h^2}{3(1 - \theta)} \leq \tau. \quad (3.57) \]

The detailed analysis of the one-dimensional problem can be found in [94].

By turning to the two-dimensional problem, we will point out that the previous method cannot be applied since the finite element block matrices of the two-dimensional problem contain some positive entries in their subdiagonals.

Let us extend the equation by some material parameters. The more general form of the two-dimensional heat conduction equation with mixed boundary conditions on the domain
\( \Omega \times (0, t_{\text{max}}) \), where \( \Omega := (0, L_x) \times (0, L_y) \), is

\[
c(x, y) \frac{\partial u}{\partial t} = \nabla (\kappa(x, y) \nabla u), \quad (x, y) \in \Omega, \ t \in (0, t_{\text{max}}),
\]

\[
u|_{\Gamma_D} = \gamma(x, y), \quad \frac{\partial u}{\partial n}|_{\Gamma_N} = 0, \ t \in [0, t_{\text{max}}), \tag{3.58}
\]

\[
u(0, x, y) = u_0(x, y), \quad (x, y) \in \Omega,
\]

where \( c \) and \( \kappa \) represent the specific heat capacity and the coefficient of the thermal conductivity, respectively. The variable \( t \) denotes time, and \( x, y \) denote the space variables. Moreover, \( \gamma(x, y) \) is the given temperature at the part of the boundary \( \Gamma_D \), which is assumed to be a non-negative real function. \( \Gamma_N = \{ \partial \Omega \mid y \neq 0 \} \) and \( \Gamma_D = \{ \partial \Omega \mid y = 0 \} \), where \( u \) is the temperature of the analyzed domain. Moreover, \( \Gamma_N \) denotes a specified part of the boundary of \( \Omega \) where Neumann boundary condition is imposed, and \( \Gamma_D \) denotes the part of the boundary where Dirichlet boundary condition is imposed to the corresponding partial differential equation. We assume that \( \Gamma_D \cup \Gamma_N = \partial \Omega \) and \( \Gamma_D \cap \Gamma_N = \emptyset \).

The material parameters (i.e., \( c \) and \( \kappa \)), are not assumed to be constant but they are bounded functions, i.e., there exist constants \( c_0, c_1, \kappa_0, \) and \( \kappa_1 \) such that

\[
0 < c_0 \leq c(x, y) \leq c_1 < \infty, \tag{3.59}
\]

\[
0 < \kappa_0 \leq \kappa(x, y) \leq \kappa_1 < \infty. \tag{3.60}
\]

The weak formulation of problem (3.58) is to find \( u \), such that

\[
\int_{\Omega} c(x, y) \frac{\partial u}{\partial t} v(x, y) dx dy + \int_{\Omega} \kappa(x, y) \nabla u \nabla v dx dy = 0 \tag{3.61}
\]

for all \( v \in H^1_D(\Omega) \) (see (1.20)). Hence, we seek such a function \( u(x, y, t) \) with \( u(x, y, t)|_{\Gamma_D} = \gamma(x, y) \), which belongs to \( H^1(\Omega) \) for all fixed \( t \), moreover, there exists \( \frac{\partial u}{\partial t} \), and it satisfies (3.61) for all \( v \in H^1_D(\Omega) \).
Let the domain \( \Omega \) be divided into \( 2n_xn_y \) triangle elements (see Fig. 3.1). We seek the spatially semi-discretized temperature \( u_h \) in the form

\[
u_h(t, x, y) = \sum_{i=0}^{n_x} \sum_{j=0}^{n_y} \phi_{i,j}(t)v_{i,j}(x, y),
\]

(3.62)

where \( v_{i,j}(x) \) are the following shape functions:
\[ v_{i,j}(x) := \begin{cases} 1 - \frac{1}{h_x}(x_i - x) - \frac{1}{h_y}(y_j - y), & \text{if } (x, y) \in \omega_{i}^{1,j} \\ 1 - \frac{1}{h_x}(x_i - x), & \text{if } (x, y) \in \omega_{i}^{2,j} \\ 1 + \frac{1}{h_x}(y_j - y), & \text{if } (x, y) \in \omega_{i}^{3,j} \\ 1 + \frac{1}{h_x}(x_i - x) + \frac{1}{h_y}(y_j - y), & \text{if } (x, y) \in \omega_{i}^{4,j} \\ 1 + \frac{1}{h_x}(x_i - x), & \text{if } (x, y) \in \omega_{i}^{5,j} \\ 0, & \text{otherwise} \end{cases} \tag{3.63} \]

\( \phi_{i,j}(t) \) are unknown (coefficient) functions for all \( i = 0, 1, \ldots, n_x \) and \( j = 1, 1, \ldots, n_y \), and \((n_x + 1) \cdot (n_y + 1)\) is the total number of nodes, and \( h_x \) and \( h_y \) are the lengths of the spatial approximations in the different directions. The unknown temperature index \( j \) runs from 1, hence, due to the boundary condition at \( y = 0 \) the temperature is known, namely, \( \phi_{0,i}(t) = \gamma \), for all \( i = 0, 1, \ldots, n_x \).

Substituting (3.62) into (3.61), we get the weak semi-discretized system of equations.

By redistributing the indices, the following equations draw up:

\[ \sum_{i=0}^{n_x} \sum_{k=1}^{n_y} \phi_{i,k}'(t) \int_{\Omega} c(x, y)v_{i,k}v_{j,l}dxdy + \sum_{i=0}^{n_x} \sum_{k=1}^{n_y} \phi_{i,k}(t) \int_{\Omega} \kappa(x, y) \nabla v_{i,k} \nabla v_{j,l}dxdy = 0, \tag{3.64} \]

\( j = 0, 1, \ldots, n_x, \ l = 1, 2, \ldots, n_y. \)

Let \( Q, M \in \mathbb{R}^{(n_x+1)^2 \times (n_y)^2} \) denote the so-called stiffness and mass matrices, respectively, with entries defined by the formulae

\[ (q_{k,l})_{i,j} = \int_{\Omega} \kappa(x, y) \nabla v_{i,k} \nabla v_{j,l}dxdy, \tag{3.65} \]

\[ (m_{k,l})_{i,j} = \int_{\Omega} c(x, y)v_{i,k}v_{j,l}dxdy. \tag{3.66} \]

To analyse the matrices of the system, let us assume for a moment that the heat capacity and the coefficient of thermal conductivity are constants. After performing the integration in (3.65) and (3.66) for the bilinear shape functions, the mass and the stiffness matrices have the form
\[ Q = \kappa \begin{bmatrix} Q_A/2 & Q_I & 0 & \ldots & 0 \\ Q_I & Q_A & Q_I & \ldots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \ldots & Q_I & Q_A & Q_I \\ 0 & \ldots & 0 & Q_I & Q_A/2 \end{bmatrix}, \quad (3.67) \]

and

\[ M = ch_x h_y \begin{bmatrix} M_0 & M_D & 0 & \ldots & 0 \\ M_D^T & M_A & M_D & \ldots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \ldots & M_D^T & M_A & M_D \\ 0 & \ldots & 0 & M_D^T & M_N \end{bmatrix}, \quad (3.68) \]

where

\[ Q_A = \frac{h_y}{h_x} \begin{bmatrix} \left(1 + \frac{h_x^2}{h_y^2}\right) & -1 & 0 & \ldots & 0 \\ -1 & 2 \left(1 + \frac{h_x^2}{h_y^2}\right) & -1 & \ldots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \ldots & -1 & 2 \left(1 + \frac{h_x^2}{h_y^2}\right) & -1 \\ 0 & \ldots & 0 & -1 & \left(1 + \frac{h_x^2}{h_y^2}\right) \end{bmatrix}, \quad (3.69) \]

\[ Q_I = \frac{h_x}{h_y} \begin{bmatrix} -1/2 & 0 & 0 & \ldots & 0 \\ 0 & -1 & 0 & \ldots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \ldots & 0 & -1 & 0 \\ 0 & \ldots & 0 & 0 & -1/2 \end{bmatrix}, \quad (3.70) \]
Then (3.64) can be expressed as

\[ M \Phi'(t) + Q \Phi(t) = 0, \quad t > 0, \quad (3.73) \]

where \( \Phi(t) \in \mathbb{R}^{(n_x+1) \times (n_y)} \) is a vector function with the components \( \phi_{i,j}(t) \). For the time discretization of the system of ordinary differential equations (3.73) we apply again the \( \theta \)-method, which results in the following scheme:

\[ M \Phi^{\ell+1} - \Phi^{\ell} = \tau Q \left( \theta \Phi^{\ell+1} + (1 - \theta) \Phi^{\ell} \right), \quad (3.74) \]

where \( \ell \) denotes the time level. It is obvious that this is a system of linear algebraic equations with respect to the unknown vector \( \Phi^{\ell+1} \) being the approximation of the temperature at the new time level \( (\ell + 1) \), which is an array on the discretization of the domain. It is worth emphasizing that the method yields the Crank-Nicolson implicit method of second order for \( \theta = 0.5 \) \[20\].

Moreover, \( h_x \) and \( h_y \) are the lengths of the spatial approximations. For the matrices of the one-dimensional linear spline functions see \[19\].

Using the Finite Element \( \theta \)-Method (3.74) to the discretization of (3.73) the following system of linear algebraic equations is obtained:

\[ X_1 \Phi^{\ell+1} = X_2 \Phi^{\ell}, \quad \ell = 0, 1, \ldots, \quad (3.75) \]

where \( X_1 = M + \tau \theta Q, X_2 = M - \tau (1 - \theta) Q \). It is clear that for all \( \Phi^{\ell+1} \) to be non-negative,
the non-negativity of the iteration matrix

\[ X = X_1^{-1}X_2 \]  

(3.76)

is required. The sufficient condition of the non-negativity of \( X \) is the following:

\[ X_1^{-1} \geq 0 \quad \text{and} \quad X_2 \geq 0. \]  

(3.77)

**Remark 3.3.4** By substracting the matrix \( X_2 \) from \( X_1 \), we get the decomposition \( X_1 - X_2 = \tau Q \) with the property (3.77), which is called a regular matrix splitting of the matrix \( \tau Q \).

For \( X_2 \) it is easy to give a condition that guarantees its non-negativity by analyzing the entries of the matrix. By a direct computation we get the condition

\[ \frac{h_y h_x c}{2 \tau} - 2 \left( \frac{h_x}{h_y} + \frac{h_y}{h_x} \right) (1 - \theta) \kappa \geq 0, \]  

(3.78)

which yields the upper bound

\[ \frac{h_x h_y c}{4 \left( \frac{h_x}{h_y} + \frac{h_y}{h_x} \right) (1 - \theta) \kappa} \geq \tau, \]  

(3.79)

assuming that \( \theta \neq 1 \), otherwise, there is no upper bound for the time step size.

By the so-called M-matrix method (i.e., finding the conditions under which \( X_1 \) is an M-matrix, [94]), it is not possible to obtain a sufficient condition for the non-negativity of the matrix \( X_1^{-1} \). This also follows from the fact that \( X_1 \) contains positive entries in its off-diagonal, therefore it cannot be an M-matrix. Hence, a sufficient condition for the inverse-positivity of matrix \( X_1 \) will be constructed by some other techniques in the following sections.

### 3.3.3 Tridiagonal Matrix Algorithm

The tridiagonal matrix algorithm (TDMA) or the Thomas algorithm (named after Llewellyn Hilleth Thomas – British physicist and applied mathematician), is a simplified form of the well-known Gaussian elimination that can be used to solve tridiagonal systems of linear algebraic equations. A scalar tridiagonal system for \( n \) scalar unknowns \( x_i \) may be written as

\[ a_i x_{i-1} + b_i x_i + c_i x_{i+1} = f_i, \quad \text{for all} \quad i = 1, \ldots, n, \]  

(3.80)
where \(a_i, b_i, c_i, d_i \in \mathbb{R}\), moreover \(a_0\) and \(c_n\) are equal to zero by definition. For such a system, the solution vector can be obtained by \(O(n)\) arithmetic operations instead of \(O(n^3)\), required by the standard Gaussian elimination \[19\]. In the literature it is sometimes referred to as the "Pendulum method", which originates from the operation of the method. Namely, first we eliminate the entries \(a_i\) of the lower subdiagonal, and then by applying a backward substitution we compute the solution \(x_i\).

By applying a finite difference or finite element method, the solution of a mathematical problem often reduces to the solution of a system of linear equations with a tridiagonal or, in higher dimensions, a block-tridiagonal matrix. If the matrix \(A\) on the left-hand side of the equation

\[
AX = F \quad (3.81)
\]

is a block-tridiagonal matrix with \((m+1) \times (m+1)\) blocks, then the problem is equivalent to the solution of the following system:

\[
B_0X_0 - C_0X_1 = F_0, \quad (3.82)
\]

\[
-A_iX_{i-1} + B_iX_i - C_iX_{i+1} = F_i, \quad i = 1, \ldots, m-1, \quad (3.83)
\]

\[
-A_mX_{m-1} + B_mX_m = F_m, \quad (3.84)
\]

where the blocks \(A_i, B_i, C_i \in \mathbb{R}^{n \times n}\); \(X_i, F_i \in \mathbb{R}^n\).

Usually, in the TDMA (tridiagonal matrix algorithm) \[86\] the solution is sought in the form

\[
X_i = \alpha_{i-1}X_{i-1} + \beta_{i-1}, \quad i = 1, 2, \ldots, m, \quad (3.85)
\]

where \(\alpha_i \in \mathbb{R}^{n \times n}\) for \(i = 0, \ldots, m-1\) and \(\beta_i \in \mathbb{R}^n\) for \(i = -1, \ldots, m-1\). Then the solution of the system of linear algebraic equations \((3.82), (3.83)\) and \(3.84\) can be defined by the following algorithm:
TDMA

1. We put

\[ \alpha_{m-1} = B_m^{-1} A_m, \quad (3.86) \]

and

\[ \beta_{m-1} = B_m^{-1} F_m. \quad (3.87) \]

2. We find \( \alpha_{m-2}, \alpha_{m-3}, \ldots, \alpha_0 \) and \( \beta_{m-2}, \beta_{m-3}, \ldots, \beta_{-1} \) by the formulas

\[ \alpha_{i-1} = (B_i - C_i \alpha_i)^{-1} A_i, \quad (3.88) \]

and

\[ \beta_{i-1} = (B_i - C_i \alpha_i)^{-1} (C_i \beta_i + F_i). \quad (3.89) \]

3. Then by the formulas

\[ X_0 = \beta_{-1} \quad (3.90) \]

and (3.85) we define \( X_0, X_1, \ldots, X_m \).

In the following, we will show under which conditions the inverses of the matrices given in the TDMA exist.

**Lemma 3.3.4** Let \( M \in \mathbb{R}^{n \times n} \). If \( \| M \| < 1 \), then \( I - M \) is regular, and

\[ \frac{1}{1 + \| M \|} \leq \| (I - M)^{-1} \| \leq \frac{1}{1 - \| M \|}. \quad (3.91) \]

For the proof see, e.g., [85].

**Lemma 3.3.5** Assume that the following conditions hold:

- there exists \( B_i^{-1} \), for all \( i = 0, 1, \ldots, m \),
- \( \| B_0^{-1} C_0 \| < 1 \),
\[ \| B_i^{-1} A_i \| + \| B_i^{-1} C_i \| < 1, \quad i = 1, \ldots, m - 1, \]
\[ \| B_m^{-1} A_m \| < 1. \]

Then
\[ \| \alpha_i \| < 1, \quad i = 0, \ldots, m - 1. \tag{3.92} \]

**Proof:** We will use mathematical induction in the course of the proof of the lemma for the values \( i = m - 1, m - 2, \ldots, 1. \)

1. For \( \alpha_{m-1} \) the statement follows form the last assumption of the lemma and (3.86).
2. Now, we will show that \( \| \alpha_i \| < 1 \) implies the relation \( \| \alpha_{i-1} \| < 1, \) for \( i = 1, \ldots, m - 2. \)

It is easy to verify the relation
\[ \alpha_{i-1} = (I - R_i)^{-1} B_i^{-1} A_i, \tag{3.93} \]
where \( R_i := B_i^{-1} C_i \alpha_i. \)

From (3.93) we get:
\[ \| \alpha_{i-1} \| \leq \| (I - R_i)^{-1} \| \| B_i^{-1} A_i \|. \tag{3.94} \]

According to the third assumption of the lemma \( \| B_i^{-1} C_i \| < 1 \) and the induction assumption, we have \( \| \alpha_i \| < 1 \) and \( \| R_i \| \equiv \| B_i^{-1} C_i \alpha_i \| < 1. \) Applying Lemma 3.3.4 we obtain
\[ \| \alpha_{i-1} \| \leq \frac{1}{1 - \| R_i \|} \| B_i^{-1} A_i \| \leq \]
\[ \leq \frac{1}{1 - \| B_i^{-1} C_i \| \| \alpha_i \|} \| B_i^{-1} A_i \| \leq \]
\[ \leq \frac{1}{1 - \| B_i^{-1} C_i \|} \| B_i^{-1} A_i \|. \tag{3.95} \]

Due to the second and third assumptions of the lemma, the right-hand side of (3.95) is less than one, therefore the lemma is proven.

**Theorem 3.3.3** Under the conditions of Lemma 3.3.5 the matrices
\[ (B_i - C_i \alpha_i)^{-1} \equiv (I - R_i)^{-1} B_i^{-1}, \quad i = 1, \ldots, m - 1. \tag{3.96} \]
exist in (3.88) and (3.89).
Proof: According to the proof of Lemma 3.3.5, \( \|R_i\| < 1 \), i.e., \((I - R_i)\) is regular, thus, it is invertible. Since we assumed that \( B_i \) is invertible, the right-hand side of (3.96) exists, and the theorem is proven. \( \blacksquare \)

Remark 3.3.5 In the algorithm (for scalar coefficients, i.e., \( n = 1 \)) of Thomas and Zhou [98], with a given \( X_0 \) the solution with \( F_i = 0 \), for \( i = 1, \ldots, m \) is sought in the form:

\[
X_i = \alpha_{i-1}X_0, \quad (3.97)
\]

where \( \alpha_{i-1} \) was calculated in a more complicated way, with introducing some unnecessary iterations as well. Clearly, we assume that \( X_0 \) is not the null matrix. Otherwise, \( X_i \equiv 0 \), for \( i = 1, 2, \ldots, m \). Substitute the above form of \( X_i \) into (3.82)–(3.84) and multiply the equations by the pseudoinverse of \( X_0 \) from the right side, then the TDMA above holds with \( \beta_i = 0 \) for all \( i \), and it is equivalent to the algorithm of [98].

In this work the authors have analyzed the minimum time step size that needs to be used in the numerical solution of diffusion problems by applying the finite element method, for the non-negativity preservation property.

Remark 3.3.6 If we seek the solution in the form

\[
X_i = \alpha_{i+1}X_{i+1} + \beta_{i+1}, \quad i = 0, \ldots, m - 1, \quad (3.98)
\]

with zero right-hand side in (3.83) and (3.84), similarly to Remark 3.3.5, it gives the same algorithm as

\[
X_i = \alpha_{i+1}X_m, \quad (3.99)
\]

if we assume that \( X_m \) is not the null matrix.

Henceforth, we analyze those conditions under which the non-negativity property of the numerical solution is preserved. Namely, we seek sufficient condition of the blocks for the following implication:

\[
F_i \geq 0 \Rightarrow X_i \geq 0, \quad i = 0, 1, \ldots, m. \quad (3.100)
\]

Theorem 3.3.4 Let \( n \geq 2 \) and we assume that the following conditions are satisfied:

1. the assumptions of Lemma 3.3.5 hold,
2. \( A_i \geq 0 \) for \( i = 1, \ldots, m \),
3. $C_i \geq 0$ for $i = 0, \ldots, m - 1$,

4. $B_i$ is monotone for $i = 0, \ldots, m$.

Then, for any $F \geq 0$, the TDMA results in a non-negative matrix $X$.

**Proof:** For $n \geq 2$ let us assume that the conditions of the theorem are satisfied. Then it is sufficient to prove that $\alpha_i \geq 0$ for $i = 0, \ldots, m - 1$ and $\beta_i \geq 0$ for $i = -1, \ldots, m - 1$. It is clear that $\alpha_{m-1}$ and $\beta_{m-1}$ are non-negative due to the assumptions 2, 3 and 4. If we show that $(B_i - C_i \alpha_i)^{-1} \geq 0$, then the proof of the theorem is complete. According to the proof of Theorem 3.3.3, $I - R_i$ is regular, and the Neumann series, $\sum_{k=0}^{\infty} R_i^k$, converges in the operator norm [100], thus, $I - R_i$ is invertible, and its inverse is the sum of the series

$$
(I - R_i)^{-1} = I + R_i + R_i^2 + R_i^3 + \ldots
$$

Theorem is proved, hence $R_i \geq 0$ for $i = 0, \ldots, m - 1$.

**Remark 3.3.7** Using the notations of (3.80), for $n = 1$ the theorem above yields the following. We assume, that the corresponding conditions are satisfied, i.e.,

1. the assumptions of Lemma 3.3.5 hold, i.e., $a_i \frac{b_i}{b_i} + c_i \frac{b_i}{b_i} < 1$ for all $i = 0, \ldots, m$,

2. $a_i \geq 0$ for $i = 1, \ldots, m$,

3. $c_i \geq 0$ for $i = 0, \ldots, m - 1$,

4. $b_i > 0$ for $i = 0, \ldots, m$.

Then for any $f \geq 0$, the scalar TDMA results in a non-negative vector $x$.

### 3.3.4 Applications of the Tridiagonal Matrix Algorithm

For various discrete one-dimensional diffusion problems the minimum time step size for the non-negativity preservation property has been studied in a similar approach by many authors, see, e.g., [81, 98]. Thomas and Zhou [98] have proposed an approach to develop a sufficient condition for the non-negativity preservation in the finite element method of one-dimensional diffusion problems, applying a backward difference time-stepping algorithm for the temporal discretization. In our earlier work [94], we pointed out its imperfections and theoretically analyzed the possible choice of the time step size and established the interval where the discrete model is reliable to the original physical phenomenon. Moreover, we extended the analysis to the $\theta$-method as well. The main result of the work is the following:
Theorem 3.3.5 Let us assume that the condition
\[
\frac{h^2 c}{6 \theta \kappa} < \tau \leq \frac{h^2 c}{3(1 - \theta) \kappa}
\]
holds. Then for the two-dimensional problem (3.1) extended by given constant material parameters \(c\) and \(\kappa\), with arbitrary non-negative initial condition the linear finite element method results in a non-negative numerical solution on any time level.

In our other work [95] the bounds of the time step size for the two-dimensional classical diffusion problem was investigated, which can be used in the finite element method applying bilinear shape functions [80] on a square domain.

Theorem 3.3.6 Let us assume that for the discretization parameters \(h\) and \(\tau\) the condition
\[
\frac{h^2 c}{12 \theta \kappa} \left(3 + \sqrt{14}\right) \leq \tau \leq \frac{h^2 c}{8(1 - \theta) \kappa}
\]
holds, where \(h\) is the length of the spatial approximation in both space directions. Then for the two-dimensional problem (3.58), with arbitrary non-negative initial condition, the bilinear finite element method results in a non-negative solution on any time level.

Earlier by the application of Theorem 3.3.4 for the one-dimensional linear finite element method [96] we got a sufficient condition for the non-negativity preservation property. However, in the sequel we will point out that in the two-dimensional bilinear finite element method we will need a different approach to give a sufficient condition for the non-negativity preservation property.

According to the notations in (3.82)–(3.84) let \(n = n_x + 1\) and \(m = n_y\). Considering the fact that the matrices of equation (3.74) are block-tridiagonal matrices, the following system can be obtained for the unknown vector \(\Phi^{\ell+1} \in \mathbb{R}^{(n_x+1) \times (n_y)}\) vector (for the sake of simplicity we will use \(\Phi\) without the superscript):

\[
\begin{align*}
-A \Phi_0 + B \Phi_1 - C \Phi_2 &= F_1, \\
-A \Phi_1 + B \Phi_2 - C \Phi_3 &= F_2, \\
&\quad \vdots \\
-A \Phi_{m-2} + B \Phi_{m-1} - C \Phi_m &= F_{m-1}, \\
-A \Phi_{m-1} + D \Phi_m &= F_m.
\end{align*}
\]
where $A, B, C, D$ are $n$-by-$n$ tridiagonal matrices and, due to the boundary condition, $\Phi_0$ is given. Since

$$A = -\left(\frac{\chi_x h_y M_D^T}{\tau} - \kappa \theta Q I\right),$$

and according to the structure of $M_D^T$, its offdiagonal contains some negative entries, i.e., the second and third assumption of Theorem 3.3.4 cannot be satisfied. However, a sufficient condition for the non-negativity preservation can be obtained in a different way.

**Lemma 3.3.6** [79] Let $A$ be an $n$-by-$n$ matrix, denote $A_d$ and $A^-$ the diagonal and the negative offdiagonal part of the matrix $A$, respectively.

Let $A^- = A^z + A^s = (a^-_{ij}) + (a^s_{ij})$. If

$$a_{ij} \leq \sum_{k=1}^{n} a^z_{ik} a^{-1}_{kk} a^s_{kj}, \text{ for all } a_{ij}, i \neq j,$$

then $A$ is a product of two M-matrices, i.e., $A$ is monotone.

We will analyse the monotonicity of $X_1$ with the help of this lemma. It is possible because it is a square matrix and it can be decomposed into the diagonal part, the positive offdiagonal part, the upper triangular and lower triangular negative parts. All the conditions of the lemma are satisfied if

$$\frac{1}{12} \leq \left(\frac{1}{12} - \frac{\kappa \tau \theta}{h_x^2 c}\right) \left(\frac{1}{12} - \frac{\kappa \tau \theta}{h_y^2 c}\right),$$

which implies the lower bound

$$\frac{h_x^2 c}{12 \theta \kappa} \left(3 \left(\frac{h_x^2}{h_y^2} + 1\right) + \sqrt{\frac{9}{4} \left(\frac{h_x^4}{h_y^4} + 1\right) + \frac{19}{2} \left(\frac{h_x^2}{h_y^2}\right)}\right) \leq \tau,$$

assuming that $\theta \neq 0$. Hence, the next statement is proven.

**Theorem 3.3.7** Let us assume that the conditions (3.79) and (3.105) hold. Then for the problem (3.58) on a rectangular domain with an arbitrary non-negative initial condition the linear finite element method results in a non-negative solution on any time level.

**Remark 3.3.8** With function coefficients instead of conditions (3.79) and (3.105) under the conditions

$$\frac{h_x h_y c_0}{4 \left(\frac{h_x}{h_y} + \frac{h_y}{h_x}\right)(1 - \theta) \kappa_1} \geq \tau,$$

60
\begin{equation}
\frac{h_y^2 c_1}{12 \theta \kappa_0} \left( \frac{3}{2} \left( \frac{h_x^2}{h_y^2} + 1 \right) + \sqrt{\frac{9}{4} \left( \frac{h_x^4}{h_y^4} + 1 \right) + \frac{19}{2} \left( \frac{h_x^2}{h_y^2} \right)} \right) \leq \tau, \tag{3.107}
\end{equation}

the non-negativity preservation property holds for the bilinear FEM solution of (3.58).

**Remark 3.3.9** If we assume that \( h_x/h_y = 1 \) (i.e., we are using the same mesh-size in both space directions), then the conditions we get for a square domain are equivalent to the conditions in (3.103).

**Remark 3.3.10** Let us introduce the notation
\begin{equation}
\sigma := \frac{3}{\left( \frac{h_y^2}{h_x^2} + 1 \right) \left( \frac{3}{2} \left( \frac{h_x^2}{h_y^2} + 1 \right) + \sqrt{\frac{9}{4} \left( \frac{h_x^4}{h_y^4} + 1 \right) + \frac{19}{2} \left( \frac{h_x^2}{h_y^2} \right)} \right)}. \tag{3.108}
\end{equation}

With function coefficients, \( \theta \) is related to the limits of the coefficient functions. Namely the conditions (3.106) and (3.107) could be satisfied only if the upper limit is greater than the lower limit for the time step size. This implies that the positivity of the right-hand side can be guaranteed by the assumption
\begin{equation}
\frac{c_1 \kappa_1}{c_0 \kappa_0} := \delta \leq \sigma \frac{\theta}{1 - \theta}, \tag{3.109}
\end{equation}

where \( c_0, c_1, \kappa_0, \kappa_1 \) were defined earlier in (3.59) and (3.60).

The above condition is equivalent to
\begin{equation}
\frac{\delta}{\sigma + \delta} \leq \theta, \tag{3.110}
\end{equation}

and this implies that if \( \delta \) monotonously tends to infinity (i.e., the amplitude of the oscillation of the functions \( c(x) \) or \( \kappa(x) \) tends to infinity), then the parameter \( \theta \), which could guarantee the non-negativity preservation property, tends monotonously to 1.

**Remark 3.3.11** If \( \theta = 1 \), there is neither any upper bound for the time step size, nor any condition for the ratio of the lengths of the spatial approximations. However, if \( \theta \) tends to zero, the lower bound tends to infinity, hence there is no such time step that can guarantee the non-negativity of the solution by our results.

**Remark 3.3.12** If the conditions of Theorem 3.3.7 hold, then we get the following complementary properties by assuming that the upper bound is always greater than the lower bound on the time step size:
Figure 3.2: The ratio of the lengths of the spatial approximations in the function of \( \theta \), with three different values of \( \delta \).

For the ratio of the lengths of the spatial approximations

\[
\sqrt{\omega - \omega^2 - 1} \leq \frac{h_x}{h_y} \leq \sqrt{\omega + \omega^2 - 1},
\]

(3.111)

where

\[
\omega = \frac{10 T^2 \delta^2 + 2T \delta^2 - 1}{-\frac{10}{9} T^2 \delta^2 - 2T \delta^2 - 1},
\]

(3.112)

and

\[
T = \frac{1 - \theta}{\theta}.
\]

(3.113)

It is worth emphasizing that the lower and the upper bound of the ratio of the lengths of the spatial approximations are reciprocal to each other (Fig. 3.2).

For \( \theta \), the parameter of the numerical method:

\[
\theta \geq \frac{1}{\sqrt{\frac{81}{100} \delta^2 + \frac{9}{20} + \frac{1}{10}}},
\]

(3.114)

for any \( \delta \), the right-hand side is approximately greater than 0.818, which implies that for the Crank-Nicolson method (\( \theta = 0.5 \)) we cannot guarantee the non-negativity by this principle [34].
Table 3.1: Time steps of the numerical models, bounds and negative results

<table>
<thead>
<tr>
<th>ID</th>
<th>Time steps</th>
<th>( \tau )</th>
<th>( \tau_- )</th>
<th>( \tau_+ )</th>
<th>( u_- )</th>
</tr>
</thead>
<tbody>
<tr>
<td>NNP1</td>
<td>30</td>
<td>1.00E-02</td>
<td>1.69E-06</td>
<td>2.50E-07</td>
<td>0</td>
</tr>
<tr>
<td>NNP2</td>
<td>100</td>
<td>1.00E-10</td>
<td>9.39E-07</td>
<td>1.25E-06</td>
<td>-5.97E-02</td>
</tr>
<tr>
<td>NNP3</td>
<td>100</td>
<td>1.00E-06</td>
<td>9.39E-07</td>
<td>1.25E-06</td>
<td>0</td>
</tr>
<tr>
<td>NNP4</td>
<td>50</td>
<td>1.00E-02</td>
<td>8.45E-04</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>NNP5</td>
<td>1</td>
<td>1.00E-02</td>
<td>6.16E-03</td>
<td>6.90E-04</td>
<td>-1.23E02</td>
</tr>
<tr>
<td>NNP6</td>
<td>1</td>
<td>2.00E-02</td>
<td>3.08E-03</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>COM</td>
<td>10</td>
<td>1.00E-04</td>
<td>3.51E-02</td>
<td>-</td>
<td>-8.42E-02</td>
</tr>
</tbody>
</table>

3.4 Numerical Experiments

In the following numerical experiments the heat conduction equation is solved on a rectangular domain by the bilinear finite element method on the mesh, illustrated in Fig. 3.1.

In the numerical experiments the TDMA was used for the inversion of the sparse tridiagonal matrices [86] and it was developed with MATLAB\textsuperscript{®}. The following figures are in three dimensions, in Fig. 3.3 and 3.6 the first axis is the time, the second one is the spatial variable \( y \) at \( x = L_x \), and the third one is temperature at the nodes.

The numerical parameters and the results for the conditions and the minimum values of the temperature during the running can be found in Table 3.1 and Table 3.2, where according to the conditions (3.78) and (3.105) we denote

\[
\tau_- := \frac{h_x^2 c}{12 \theta \kappa} \left( \frac{3}{2} \left( \frac{h_x^2}{h_y^2} + 1 \right) + \sqrt{\frac{9}{4} \left( \frac{h_x^4}{h_y^4} + 1 \right) + \frac{19}{2} \left( \frac{h_x^2}{h_y^2} \right)^2} \right),
\]

\[
\tau_+ := \frac{h_x h_y c}{4 \left( \frac{h_x}{h_y} + \frac{h_y}{h_x} \right) (1 - \theta) \kappa},
\]

moreover

\[
u_- = \min_{(t,x,y)} (u(t,x,y)), \quad (x,y) \in \Omega, \quad t \in [0,t_{\text{max}}],
\]

where the variables \( t,x,y \) are considered in discrete sense, based on the numerical parameters of the experiments.

First, we apply the Crank-Nicolson scheme and a relatively long time step (see Table 3.1 NNP1), which results in a negative \( X_2 \). In Fig. 3.3, however, the results are non-negative, one can see that the numerical method is quite unstable, hence there is an oscillation with decreasing tendency in the solution. Moreover, there are some grid points
Table 3.2: Numerical parameter of the method, spatial and material parameters, boundary condition at the given side of the domain

<table>
<thead>
<tr>
<th>ID</th>
<th>$\theta$</th>
<th>$h_x$</th>
<th>$h_y$</th>
<th>$n$</th>
<th>$m$</th>
<th>$c$</th>
<th>$\kappa$</th>
<th>$\gamma$</th>
<th>$u_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>NNP1</td>
<td>0.5</td>
<td>2.50E-03</td>
<td>5.00E-03</td>
<td>20</td>
<td>50</td>
<td>1</td>
<td>10</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>NNP2</td>
<td>0.9</td>
<td>2.50E-03</td>
<td>5.00E-03</td>
<td>20</td>
<td>50</td>
<td>1</td>
<td>10</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>NNP3</td>
<td>0.9</td>
<td>2.50E-03</td>
<td>5.00E-03</td>
<td>20</td>
<td>50</td>
<td>1</td>
<td>10</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>NNP4</td>
<td>1</td>
<td>2.50E-02</td>
<td>5.00E-02</td>
<td>20</td>
<td>30</td>
<td>1</td>
<td>1</td>
<td>100(1 + sin(8\pi t))</td>
<td>0</td>
</tr>
<tr>
<td>NNP5</td>
<td>0.5</td>
<td>1.00E-01</td>
<td>4.00E-02</td>
<td>18</td>
<td>23</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>300</td>
</tr>
<tr>
<td>NNP6</td>
<td>1</td>
<td>1.00E-01</td>
<td>4.00E-02</td>
<td>18</td>
<td>23</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>300</td>
</tr>
<tr>
<td>COM</td>
<td>1</td>
<td>2.50E-01</td>
<td>2.50E-01</td>
<td>5</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Figure 3.3: The solution obtained by the Crank-Nicolson scheme and for relatively large time step (NNP1)

where the temperature is higher than the temperature of the source, which is physically impossible according to the conservation of energy. It is worth emphasizing that, in this case, according to our results, there is no such time step which could guarantee the non-negative solution of the problem ($\tau_- > \tau_+$).

When we apply smaller time steps than those according to (3.105) (NNP2), then there will be small negative peaks close to the first node (magnified part in Fig. 3.4). These solutions are unrealistic, since the absolute temperature should be non-negative.

For the sake of completeness, in Fig. 3.5 we applied the time step size from the interval (3.79) and (3.105) (NNP3), and it can be seen that we have got a more stable numerical method. In this figure the first two-dimensions are the spatial ones ($x, y$) and the third is the temperature at the nodes. It is easy to see that, by use of appropriate time steps,
Figure 3.4: The solution obtained by too small time step (NNP2)

the solution becomes much smoother than in Fig. 3.3 or in Fig. 3.4 .

It is worth emphasizing that the time dependency of the Dirichlet boundary condition does not affect our analysis. In Fig. 3.6 we applied the time step size from the interval (3.79) and (3.105) (NNP4), and it can be seen that we have got a stable numerical method, and the changes in the boundary conditions are appearing continuously inside the domain by time.

Remark 3.4.1 It can be easily shown, that if instead of the given mixed boundary conditions homogeneous Dirichlet boundary conditions are applied to problem (3.58), the same conditions can be obtained for the non-negativity preservation property.

In Fig. 3.7 and Fig. 3.8 the results were presented applying homogeneous Dirichlet boundary condition to the two-dimensional problem (3.58). We got similar results to the previous case, i.e., by using the Crank-Nicolson method (Fig. 3.7 (NNP5)) we cannot guarantee the non-negativity of the solution. In Fig. 3.8 the solution is preserving its sign as it was expected, since here we applied a time step from the obtained interval (3.79) and (3.105) (NNP6) .

COMSOL Multiphysics 3.51a (formerly FEMLAB) is a finite element method solver and simulation software package for various physics and engineering applications, especially for coupled phenomena, or multiphysics. The researchers and engineers usually using COMSOL Multiphysics like softwares as a black box, i.e., they don not exactly know the numerical parameters of their model, neither those of the numerical method.
Figure 3.5: The solution obtained by applying a time step from the interval (3.79) and (3.105) (NNP3)

Figure 3.6: The solution obtained by applying a time step from the interval (3.79) and (3.105) with time-dependent Dirichlet boundary condition (NNP4)
Figure 3.7: The solution obtained by applying Cranck-Nicolson method and homogeneous Dirichlet boundary condition \((NNP5)\)

Figure 3.8: The solution obtained by applying a time step from the given interval and homogeneous Dirichlet boundary condition \((NNP6)\)
Figure 3.9: Cross-section plot of the false solution obtained by COMSOL Multiphysics (COM)

Our numerical experiment in COMSOL Multiphysics has shown that this attitude could cause false results. We have solved the two-dimensional heat conduction equation (COM), with time step outside of the interval $[3.79]$ and $[3.105]$ (Fig. 3.9). This solution is false, since the absolute temperature measured in kelvin should be non-negative.

### 3.5 On Relations Between Qualitative Properties

Previously we have shown that several problems for partial differential equations possess some characteristic qualitative properties which are typical of the phenomenon the partial differential equation describes. The most important two of them are the maximum principles (for elliptic equations see Chapter 2) and the non-negativity preservation property (for parabolic equations see Chapter 3).

The relations and the implications between these qualitative properties have been revealed only recently by Faragó and Horváth [29]. They have analyzed the connections between the different qualitative properties of numerical solutions of linear parabolic problems with Dirichlet type boundary condition. For the detailed analysis see, e.g., [43], [30] for parabolic problems, and [63], [70] for elliptic problems.

In this short section our aim is to give some information about the relation between the obtained non-negativity preservation property and the maximum principles for parabolic and elliptic partial differential equations.

Let us consider the classical solution of the problem (2.9)–(2.10). Then the non-negativity preservation property reads as follows.

\[ f(x) \geq 0 \text{ , for all } x \in \Omega \Rightarrow u(x) \geq 0, \text{ for all } x \in \Omega. \]  

(3.118)
Remark 3.5.1 According to the assumptions of the material parameters $c$ and $\delta$ it is easy to see, that if $f(x) \geq 0$, for all $x \in \overline{\Omega}$, then Theorem 2.3.1 is valid by writing zero in the place of the given lower estimation, which yields exactly the non-negativity preservation property. Considering this, it is easy to see that the maximum principles imply the non-negativity preservation for the discrete results as well.

Now to obtain the maximum principle from the non-negativity preservation property for the parabolic problem presented in Chapter 3 let us consider the linear finite element discretization (3.74) of the given two-dimensional parabolic problem. Using its notation, we have the following result.

Theorem 3.5.1 Assume that condition (3.32) (which guarantees the non-negativity of the solution of (3.26)) holds and $X_1$ is SDD and monotone. Then the following estimation is valid for any time level $\ell$:

$$0 \leq y^{\ell+1} \leq \max \left\{ 0, \max_{j=1, \ldots, n} \frac{X_2 y^{\ell+1}}{\alpha_j(X_1)} \right\}.$$  \hspace{1cm} (3.119)

Proof: Let us use the notation of Theorem 2.4.1 i.e., denote $A = X_1$, $F = X_2 y^{\ell}$ and $u = y^{\ell+1}$ for any $\ell$ time level. The lower estimate is obvious, hence the conditions which guarantee the non-negativity of the solution holds. Since, the matrix $X_1$ is assumed to be SDD and monotone if (3.105) holds, the statements of Theorem 2.4.1 are valid, hence, the statement is proven. \hfill \blacksquare

Remark 3.5.2 Considering the structure of the matrices ($X_1$ is SDD and M-matrix, i.e., monotone) for the finite difference method in (3.26) and (3.28) and the well-known inequality

$$\max(X_2 y^{\ell}) = \|X_2 y^{\ell}\|_{\infty} \leq \|X_2\|_{\infty} \|y^{\ell}\|_{\infty},$$

since $\alpha_j(X_1) = \frac{1}{\tau}$ and $\|X_2\|_{\infty} = \frac{1}{\tau}$, the theorem above yields the following:

$$0 \leq y^{\ell+1} \leq \max y^{\ell},$$  \hspace{1cm} (3.120)

which implies the stability in the maximum norm of the system (3.26). For matrices (3.30) and (3.31) of the finite element method, we can obtain the same inequality if we assume that the off-diagonal entries of $X_1$ are non-positive.
Chapter 4

Proton Exchange Membrane Fuel Cells

Nowadays, electrical energy is the cleanest and most versatile energy that can be used in almost all fields of life. Due to the technical improvements, the utilization and the efficiency of producing electrical energy are increasingly growing. However, this technical advancement has caused huge destruction in the environment and a drastic decrease in the energy sources towards the sustenance of comfort.

One of the greatest challenges of the 21st century is not just to produce the required energy, but also to transfer it to the consumers. As a matter of fact, the required energy can be produced in different power plants. The main problem is the storage and the transfer of the produced energy. As there is no way to lay wires and pipes everywhere, the energy must be somehow “packed in”. One of the alternatives is to “close” the energy in chemical compounds, and to liberate it at the consumers when needed. Currently two versions of this method are known. Burning fossil fuels (coal, natural gas, etc.), which is polluting and not renewable, or applying fuel cells\(^1\), which is environmentally sound and partly renewable, therefore the required fuel can be mass produced artificially (hydrogen, methanol, etc.).

With the help of fuel cells, green energy can be obtained by electrochemical “burning” of the compounds (mostly organic hydride compounds, or pure hydrogen) in the right time. The technology itself is not new. The first fuel cell was made by Sir William Robert Grove in 1839. Even, the technology was stable in the industrial practice, its performance was quite low. Since then many scientists have been working on increasing the performance of fuel cells, but up to the present time nobody has been able to build a fuel cell with the efficiency of an Otto engine\(^1\).

Fuel cells are galvanic batteries that are able to convert the chemical energy of the fuel directly to electrical energy. However, one of the biggest differences between fuel cells and

\(^1\)Recently there has been some confusion concerning the correct translation of “fuel cell” into Hungarian. Many are using the word “üzemanyagcella”. This refers to the power source of vehicles, and it is not justified to use it in a broader sense. The official Hungarian chemical name is “tüzelőanyag-elem”, which just as well expresses the other, e.g., industrial and mobile applications.
galvanic batteries is the fact that while the galvanic batteries need changing or charging, the fuel cells can operate constantly by continuous reloading of its fuel.

In Fig. 4.1 the schematic structure of a proton exchange membrane fuel cell (PEMFC) is shown. Fuel cells “burn” hydrogen fuel ($H_2$) and oxygen ($O_2$) to water, but instead of heat, electrical energy is produced. This can happen if the overall chemical reaction is separated by a membrane to the oxidation and reduction. The oxidation takes place at the anode (negative electrode) and the reduction takes place at the cathode (positive electrode). Electrons pass through the circuit and generate power while protons pass through the membrane $[32]:$

\[
\begin{align*}
H_2 + \frac{1}{2}O_2 &\rightarrow H_2O \\ (\text{Overall}) \\
H_2 &\rightarrow 2H^++2e^- \\ (\text{Anode}) \\
\frac{1}{2}O_2 + 2e^- + 2H^+ &\rightarrow H_2O \\ (\text{Cathode})
\end{align*}
\]

The electro-chemical reaction takes place on the boundary of two phases (solid and solution) $[42]$. This means that the surface of the electrode and the rate of the reaction determine the amount of material that reacts in unit time and the power of the cell. Catalysts enhance the efficiency by decreasing the activation energy of the reaction, while porous electrodes enhance the active surface by several orders of magnitude. Such elec-
Electrodes are like sponge, they have plenty of inner cavities, and so an enormous active surface, where the oxidation of hydrogen and the reduction of oxygen can take place. There is an electron flow in the solid phase, and an ion flow in the solution phase. On the boundary of the two phases the electro-chemical reaction takes place, while the charge neutrality is macroscopically preserved.

Since the term "solution phase" could be very confusing, it is worth emphasizing that in our case it does not mean that there is any liquid compound in the presented type of fuel cell, but it refers only to the property of the given material.

4.1 History of Fuel Cells

In 1800, Alessandro Volta designed a device that was able to constantly produce electric current. In the same year Nicholson and Carlisle performed the first experiment to decompose water with a Volta column, in which electric current was used to produce hydrogen and oxygen from water. What we are interested in is the reverse process, i.e., the combination of hydrogen and oxygen to form water [12]. This is called exploding gas reaction, since it takes place very fiercely, with a big energy release, only by a certain $\text{H}_2/\text{O}_2$ proportion ($>2$) and at a temperature higher than 600 $^\circ\text{C}$. In the presence of platinum catalyst, $\text{H}_2$ and $\text{O}_2$ mix explosively.

Sir William Robert Grove, considered as the father of fuel cells, realized that the above reaction can be used for energy production in a galvanic battery with high efficiency even at room temperature. In 1838, during the electrolysis of water he noticed that after switching off the electric current, a current begins to flow in the opposite direction. The reason for this current is the following. The hydrogen produced on one of the platinum electrodes is oxidized, while the oxygen on the other electrode is reduced. On the base of this discovery, Grove constructed the first fuel cell, which he called gas battery to distinguish it from other batteries in which a reaction between metals and metal compounds produced the current. A gas battery consists of two platinum electrodes with one end of each immersed in sulfuric acid. The other ends separately sealed in containers of oxygen and hydrogen. Grove noticed that the level of the solutions rises when current flows between the two electrodes. This indicated the consumption of the hydrogen and oxygen. Grove wanted to create a competitor for the steam engine, but without success, at least in his life.

It is written that for more than 100 years after Grove nothing happened in the respect of the utilization of fuel cells. This is partly true. Indeed, this significant invention was not exploited, but there were continuous attempts to do so. Ostwald theoretically explained...
the processes taking place in fuel cells (1893), while the exploding gas battery of Mond and Langer (1889), as well as the carbon/air battery of Jacques (1890) proved to be operable. In Germany, Siemens was dealing with the electrochemical process called “cold burning” (hydrogen-oxygen cells), which he wanted to use mainly for the energy supply of submarines. However, the design of an efficient, high current density device only became possible when the laws of the kinetics of electrode processes have been revealed, the research of catalysts began to boom and appropriate electrolytes were constructed.

The early story ends with the research of Francis Thomas Bacon, which started in the 1930’s and opened the way for the modern development. Bacon constructed the first alkali fuel cell, which served on the Apollo spacecraft after 25 years of developing. In 1959, Bacon presented a 5 kW device [4] and in the same year Ihrig built a 20 horsepower (15 kW) fuel-cell driven tractor for the well-known american company Allis-Chalmers [55].

Certainly, the different types of fuel cells have their own stories. Phosphorous acid was for long ignored because it does not conduct electricity so well as sulfuric acid. In 1961, Elmore and Tanner realized that phosphorous acid became a satisfactory conductor at a higher temperature, moreover, as opposed to sulfuric acid, it was not reduced [24]. The first 5kW fuel cell with phosphorous acid had been prepared by 1965 for the American army, and since then the development has been unbroken. Today 50-100 kW fuel cells are used as energy sources of buses, just as their more powerful versions in the lighting and heating of buildings.

At each cell type the continuously developed new materials have always played a big role, however, the construction of polymer-electrolyte membranes was really a milestone, and a great example to demonstrate that a new material or idea can lead to a paradigm shift in some field. Polymer-electrolyte cells have been developed for spacecrafts, and only later was the technology applied for “terrestrial” use, in power plants and cars.

By the end of the 60’s, the fundamental types of fuel cell batteries had been ready to conquer the world. The traditional fuel cell elements had already been involved in foreign and Hungarian text books, however, in the 60’s modern cell types were described as well. From that on, a series of special books were published in this topic. After the Second World War, an exceptional financial support was received by the research institutes of NASA. The major aim of the Americans was to maintain and enhance their technical dominance over the USSR. Shooting the first satellite in history (Sputnik) in 1957 motivated the technological development. The most spectacular scene of the competition was space research with the aim of conquering the space. Fuel cells played an important role in the Gemini program (proton exchange membrane cells) and the Apollo program (alkali electrolyte cells).
4.2 Modeling of Fuel Cells

The importance of constructing mathematical models for fuel cells is threefold. First, it leads to a better understanding of the underlying phenomena. Second, it provides a useful tool for the optimization of fuel cell systems. Third, it will be crucial to control fuel cell based applications (vehicle, backup power, etc.) in the future. The general method to build up a reliable model starts with the selection of the phenomena that primarily influence the performance of the fuel cells under interest. The subsequent step is the description of these processes in terms of differential and/or algebraic equations, and finally the choice of an adequate mathematical scheme. The use of an analytical method leads to an exact description, however, it can be applied only for very simplified cases. A broadly used empirical model was published by Kim et al. [62], which fits the experimental curves excellently, but without the detailed interpretation of the parameters used. This model can be applied for the steady state behavior of fuel cells [92] and for real-time simulation, but it is unsuitable for the optimization of the parameters, like platinum loading (catalyst), Nafion® content (see Section 1.3), etc. The transient behavior cannot be elucidated by this model, either.

Gomadam et al. [45] studied the transient behavior of porous electrodes by using a linearized form of the kinetic equations related to the electrochemical reactions. The double-layer effect was also taken into account (see Section 1.3). The respective set of partial differential equations has been solved analytically. Based on the results of the calculation, the corresponding impedance spectra were derived for different configurations of the measurement. Because of using linear relationship, the steady-state performance in the whole current range cannot be interpreted.

Kulikovsky [73] gave two asymptotic solutions for high and low currents, respectively, assuming ideal transport of the reactants. By the expressions derived for voltage-current curves, the appearance of the double Tafel slope was elucidated. Other asymptotic solutions were given by Jaouen et al. [56] for the voltage-current curves considering the structure of the electrode, by using a spherical agglomerate model. Deeper insight and understanding of the effects of the material parameters for fuel cell performance can be achieved by the exact expressions, however, the detailed and transient behaviors cannot be predicted by analytical models. Results obtained by numerical models are less general, but both the steady-state voltage-current curves (V-I curves) and transient effects can be interpreted. These methods can be used for parameter estimation of the transport and electrochemical properties [54], and they are also useful for revealing the effects of new parameters [82].
Complex models \cite{107} are needed to solve different phenomenological equations such as the Nernst-Planck equation for multiple mass transport, the Stefan-Maxwell equation for heat transfer, Ohm’s law for ionic migration and electron conductivity, and the equations of electrochemical kinetics. These transport and transfer processes are coupled, and the equations are often highly non-linear. In practical systems (using real parameters) these processes have different time scales, as well. Moreover, each of the subsystems requires different numerical schemes with different time and spatial discretizations. These models are usually solved by using only a single numerical treatment, e.g., Runge-Kutta, Newton or Cranck-Nicholson methods. The solution of the respective numerical scheme is quite slow (sometimes slower than an experiment).

Subramanian et al. \cite{93} developed a method to reduce the number of the governing equations of Li-ion battery simulation by using different mathematical techniques. The original problem with a proper discretization has 4800 equations, which can be reduced to 49, and finally the simulation time of the discharge curve can be cut to 85 ms. However, in this model the double-layer capacitance was not included. It is quite evident that there is no such method which fulfills all of the requirements. Therefore, in this chapter of the thesis different numerical methods, such as the operator splitting techniques \cite{53}, \cite{66} are applied for the simulation of fuel cells. The procedure applied herein possesses practically all of the advantages of the other techniques mentioned previously, i.e., dealing with complexity, using different numerical schemes and its characteristic properties.

### 4.3 Governing Equations

In practice a consumer (an electric device) is inserted into the outer circuit (see Fig. 4.1), which is fed by the current arising from the fuel cell. We assume that the current in the outer circuit is known and we can control it. The aim of the following investigation is to calculate the corresponding voltage, which is called the cell potential. This gives also the electric energy provided by the fuel cell, which is very important in the course of evaluating the performance of a fuel cell.

According to Kirchoff’s law, the cell potential $E_{\text{cell}}$ can be calculated by the following equation, see also \cite{76}:

$$E_{\text{cell}}(t) = E_{\text{OC}}(t) - \eta^a(t) - \frac{W_{\text{mem}}}{\kappa_{\text{mem}}} I(t) - V^*(t), \quad (4.1)$$
<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>Specific interfacial area</td>
<td>cm(^{-1})</td>
</tr>
<tr>
<td>(C_{dl})</td>
<td>Double-layer capacitance</td>
<td>F/cm(^2)</td>
</tr>
<tr>
<td>(D)</td>
<td>Oxygen diffusion coefficient at the gas phase</td>
<td>cm(^2)/s</td>
</tr>
<tr>
<td>(D_2)</td>
<td>Oxygen diffusion coefficient at the solution phase</td>
<td>cm(^2)/s</td>
</tr>
<tr>
<td>(E_{cell})</td>
<td>Cell potential</td>
<td>V</td>
</tr>
<tr>
<td>(E_{OC})</td>
<td>Open circuit potential</td>
<td>V</td>
</tr>
<tr>
<td>(F)</td>
<td>Faraday constant (96487)</td>
<td>C/mol</td>
</tr>
<tr>
<td>(I)</td>
<td>Total current density</td>
<td>A/cm(^2)</td>
</tr>
<tr>
<td>(i_0)</td>
<td>Exchange current density at the cathode</td>
<td>A/cm(^2)</td>
</tr>
<tr>
<td>(i_0^a)</td>
<td>Exchange current density at the anode</td>
<td>A/cm(^2)</td>
</tr>
<tr>
<td>(i_1)</td>
<td>Solid phase current density at the cathode</td>
<td>A/cm(^2)</td>
</tr>
<tr>
<td>(i_2)</td>
<td>Solution phase current density at the cathode</td>
<td>A/cm(^2)</td>
</tr>
<tr>
<td>(i_f)</td>
<td>Faradaic current density</td>
<td>A/cm(^3)</td>
</tr>
<tr>
<td>(j_d)</td>
<td>Limiting current at the cathode</td>
<td>A/cm(^2)</td>
</tr>
<tr>
<td>(L)</td>
<td>Thickness of the cathode</td>
<td>cm</td>
</tr>
<tr>
<td>(L_2)</td>
<td>Characteristic size of the agglomerates</td>
<td>cm</td>
</tr>
<tr>
<td>([\text{O}_2])</td>
<td>Oxygen concentration at the cathode</td>
<td>mol/cm(^3)</td>
</tr>
<tr>
<td>([\text{O}_2]_L)</td>
<td>Oxygen concentration of the source</td>
<td>mol/cm(^3)</td>
</tr>
<tr>
<td>([\text{O}<em>2]</em>{ref})</td>
<td>Reference oxygen concentration</td>
<td>mol/cm(^3)</td>
</tr>
<tr>
<td>(R)</td>
<td>Universal gas constant (8.3144)</td>
<td>J/molK</td>
</tr>
<tr>
<td>(T)</td>
<td>Cell temperature</td>
<td>K</td>
</tr>
<tr>
<td>(u)</td>
<td>Dimensionless overpotential at the cathode</td>
<td></td>
</tr>
<tr>
<td>(V^*)</td>
<td>Potential loss at the cathode</td>
<td>V</td>
</tr>
<tr>
<td>(W_{mem})</td>
<td>Membrane thickness</td>
<td>cm</td>
</tr>
<tr>
<td>(\alpha)</td>
<td>Transfer coefficient in the cathode</td>
<td></td>
</tr>
<tr>
<td>(\alpha_a)</td>
<td>Anodic transfer coefficient at the anode</td>
<td></td>
</tr>
<tr>
<td>(\alpha_c)</td>
<td>Cathodic transfer coefficient at the anode</td>
<td></td>
</tr>
<tr>
<td>(\eta)</td>
<td>Overpotential at the cathode</td>
<td>V</td>
</tr>
<tr>
<td>(\eta^a)</td>
<td>Overpotential at the anode</td>
<td>V</td>
</tr>
<tr>
<td>(\gamma)</td>
<td>Reaction order</td>
<td></td>
</tr>
<tr>
<td>(\nu^2)</td>
<td>Dimensionless Exchange current density</td>
<td></td>
</tr>
<tr>
<td>(\phi_1)</td>
<td>Solid phase potential</td>
<td>V</td>
</tr>
<tr>
<td>(\phi_2)</td>
<td>Solution phase potential</td>
<td>V</td>
</tr>
<tr>
<td>(\kappa_{eff})</td>
<td>Effective solution phase conductivity</td>
<td>S/cm</td>
</tr>
<tr>
<td>(\sigma_{eff})</td>
<td>Effective solid phase conductivity</td>
<td>S/cm</td>
</tr>
<tr>
<td>(\sigma_{mem})</td>
<td>Membrane conductivity</td>
<td>S/cm</td>
</tr>
</tbody>
</table>

Table 4.1: List of symbols
where $t \in (0, t_{\text{max}})$ denotes time. Here $E_{\text{OC}} \approx 1.23\,\text{V}$ denotes the open circuit potential, which is present between the anode and cathode without the presence of any consumer (see Section 1.3). To solve the equation above we have to determine the potential losses at the particular parts of the fuel cell (anode, membrane, cathode) at a given load level. Since at the anode the exchange current density of the hydrogen oxidation is 3-4 orders of magnitude higher than the exchange current density at the cathode, the calculation of the potential loss at the cathode could be simplified to the solution of the following non-linear equation, however, it is worth emphasizing that the model of the cathode can be applied to the anode as well, by using different kinetics and material parameters:

$$I(t) = i_a^0(t) \left[ \exp \left( \frac{\alpha_a^a F \eta_a^a(t)}{RT} \right) - \exp \left( -\frac{\alpha_a^c F \eta_a^a(t)}{RT} \right) \right], \quad (4.2)$$

where $i_a^0(t)$ and $T$ yield the density of the exchange current at the anode and the temperature of the cell, respectively, moreover the left-hand side $I(t)$ refers to the total current density of the cell, i.e., the current in the outer circuit. The explanation for the remaining material coefficients are summarized in Table 4.1. Considering the simplest form of Ohm’s law, the term $\frac{W_{\text{mem}}}{\kappa_{\text{mem}}} I(t)$ means the potential loss at the membrane, the thickness and conductivity of which are denoted by $W_{\text{mem}}$ and $\kappa_{\text{mem}}$, respectively.

The calculation of the last quantity on the right-hand side ($V^*$), which refers to the potential loss at the cathode, needs a detailed analysis. The interval $(0, L)$ refers to the thickness of the cathode, where two phases are distinguished (see Figure 4.1):

- The solution phase, where the hydrogen ions are conducted according to the rate $\kappa_{\text{eff}}$. The potential and the current density in this phase are denoted by $\phi_2$ and $i_2$, respectively.
- In the solid phase of the cathode electrons are conducted according to the rate $\sigma_{\text{eff}}$. The potential and the current density here are denoted by $\phi_1$ and $i_1$, respectively.

All of these quantities could be allowed to depend on time and space corresponding to the given assumptions and the structure of the fuel cell and the time evolution of the process. In what follows we will use the notation $\partial_t$ and $\partial_{\xi, \xi}$ for the first and second partial derivative of the given argument function with respect to its variable $\xi$, respectively. Using the defined quantities, $V^*$ in (4.1) can be given as

$$V^*(t) = \phi_1(t, L) - \phi_2(t, 0), \quad t \in (0, t_{\text{max}}). \quad (4.3)$$
The quantity we investigate in the governing equations is the overpotential
\[ \eta(t, x) = \phi_1(t, x) - \phi_2(t, x) \geq 0, \quad x \in (0, L), \quad t \in (0, t_{\text{max}}). \] (4.4)

In the calculation of the potentials, we choose the reference level to be at the left end of the solution phase, i.e., we define \( \phi_2(t, 0) = 0 \). This is in a good accordance with the uniqueness of the solutions in the corresponding equations. As we will see, the governing equations depend only on the spatial derivatives of the potentials, such that the above assumption is necessary to determine both \( \phi_2(t, x) \) and \( \eta(t, x) \). Then an immediate consequence of (4.3) and (4.4) is that
\[ V^*(t) = \phi_1(t, L) = \eta(t, L) + \phi_2(t, L). \] (4.5)

Applying Ohm’s law for both phases we obtain
\[ i_1(t, x) = -\sigma_{\text{eff}}(x) \partial_x \phi_1(t, x), \]
\[ i_2(t, x) = -\kappa_{\text{eff}}(x) \partial_x \phi_2(t, x), \] (4.6)

and the principle of the electroneutrality gives
\[ -\partial_x i_1(t, x) = \partial_x i_2(t, x). \] (4.7)

The conservation law for the currents (see [84]) results in the formula
\[ \partial_x (\kappa_{\text{eff}}(x) \partial_x \phi_2(t, x)) = -a(x)C_{\text{dl}}(x) \partial_t \eta(t, x) - a(x)i_0(x)g\left(\frac{\alpha F}{RT} \eta(t, x)\right). \] (4.8)

Here, the function \( C_{\text{dl}}(x) \) gives the double-layer capacitance at the cathode side, and the last term yields the faradic current with \( i_0(x) \), the exchange current density at the cathode (see Section 1.3). For the notations of the material coefficients we refer to Table 4.1. The function \( g: \mathbb{R} \rightarrow \mathbb{R} \) refers to the kinetics of the oxygen reduction reaction here. This should be an increasing function with \( g(0) = 0 \).

**Remark 4.3.1** Among the several approaches for the sake of simplicity we apply linear kinetics and, accordingly, we use
\[ g_L(u) = c(x)u, \] (4.9)

where \( c(x) \) is a given non-negative function. Other possible choices are the following, which are going to be used in the course of the analysis and the numerical experiments
PROTON EXCHANGE MEMBRANE FUEL CELLS

Figure 4.2: Comparison of the different kinetics \( j_D(x) \equiv 2, \ c(x) \equiv 0.1 \)

- **Butler-Volmer kinetics:**
  \[
  g_{BV}(u) = c(x)(\exp(u) - \exp(-u)).
  \]

- **diffusion kinetics:**
  \[
  g_D(u) = j_D(x) \left( \frac{c(x)\exp(u)}{c(x)\exp(u) + j_D(x)} - \frac{c(x)\exp(-u)}{c(x)\exp(-u) + j_D(x)} \right),
  \]
  where \( j_D(x) \) is the limiting current (see Section 1.3), which in this equation is acting as a diffusion coefficient. This choice provides the most accurate model of the cathode reaction.

In what follows the notation \( g(u) \) stands for any of the above functions \( (g_L, g_{BV}, g_D) \).

**Remark 4.3.2** In what follows we assume that \( j_D(x) \) and \( c(x) \) are such functions that \( g_L(u_1) < g_D(u_2) < g_{BV}(u_3) \) holds for any functions \( u_1, u_2, u_3 \) for which \( 0 \leq u_k(x) < C \) for all \( x \in (0, L) \), \( k = 1, 2, 3 \), where \( C \) is a real constant for any given functions \( j_D(x) \) and \( c(x) \).

At the left end of the cathode only the protons can exit to the membrane and similarly, at the right end (at the current collector), only the electrons can leave the cathode. Therefore \( \partial_x \phi_1(t, 0) = 0 \) and \( \partial_x \phi_2(t, L) = 0 \) such that using (4.4) we have the following boundary conditions
\[ \begin{align*}
\partial_x \eta(t, 0) &= -\partial_x \phi_2(t, 0) = -\frac{1}{\kappa_{\text{eff}}(0)} I(t), \quad t \in (0, t_{\text{max}}), \\
\partial_x \eta(t, L) &= \partial_x \phi_1(t, L) = \frac{1}{\sigma_{\text{eff}}(L)} I(t), \quad t \in (0, t_{\text{max}}).
\end{align*} \] (4.12)

Although we have listed all physical principles and the governing equations here, the corresponding equations are not yet ready for the solution, since (4.8) contains also the unknown term \( \phi_2(t, x) \).

In the next sections we present a macro-homogeneous and a heterogeneous approach to solve the drawn system.

### 4.4 Macro-Homogeneous Model

Using the definitions and notations of the previous section, in this section we considering the macro-homogeneous model of Subramanian et al. [93], i.e.,

- only a one-dimensional model is considered,
- the concentration distributions of the gases are neglected,
- double layer charging as well as faradaic reaction occur,
- the double layer capacitance as well as all the material and kinetic parameters and properties are assumed to be contants.

It can be shown [32], that the continuous mathematical model of the porous electrode (4.5)–(4.8) under these assumptions can be transformed into the following canonical form with homogeneous initial conditions:

\[ \partial_\xi u(\delta, \xi) = \partial_{\xi\xi} u(\delta, \xi) - \nu^2 g(u(\delta, \xi)), \quad \xi \in (0, 1), \quad \delta \in (0, \delta_{\text{max}}), \] (4.13)

where \( u(\delta, \xi) := \frac{\alpha F}{RT} \eta(t, \xi) \) is the new unknown, the dimensionless overpotential and

\[ \nu^2 = \alpha \iota_0 L^2 \frac{\alpha F}{RT} \left( \frac{1}{\kappa_{\text{eff}}} + \frac{1}{\sigma_{\text{eff}}} \right) \] (4.14)

is the dimensionless exchange current density. Moreover, \( \xi \) and \( \delta \) are the new space and time variables, respectively, which are defined as

\[ \xi := \frac{x}{L}, \quad \text{and} \quad \delta := \frac{t}{p}, \] (4.15)
where
\[ p = \frac{C_{\text{dl}}RT}{i_0\alpha F}v^2. \] (4.16)

The boundary conditions (4.12) are transformed as well:
\[ \partial_\xi \eta(\delta, 0) = -\partial_\xi \phi_2(\delta, 0) = -\frac{FL}{\kappa_{\text{eff}}RT}I(\delta), \quad \delta \in (0, \delta_{\text{max}}), \]
\[ \partial_\xi \eta(\delta, 1) = \partial_\xi \phi_1(\delta, 1) = \frac{FL}{\sigma_{\text{eff}}RT}I(\delta), \quad \delta \in (0, \delta_{\text{max}}). \] (4.17)

After solving the canonical form (4.13) of problem (4.5)–(4.8) under the given assumptions, by some basic transformation one can easily get the solution of the original problem. First, let us apply the finite difference method for the spatial discretization and the \( \theta \)-method (see Section 3.3) for the time discretization of problem (4.13), then the equation
\[ X_1 u^{\ell+1} = X_2 u^\ell + f(u^\ell), \quad \ell = 0, \ldots \] (4.18)
presents the fully discretized problem, where
\[
X_1 = \begin{bmatrix}
\frac{1}{h^2} & -\frac{1}{h^2} & 0 & \ldots & 0 \\
-\frac{2\theta}{h^2} & \frac{1}{\tau} - \frac{1}{h^2} & -\frac{\theta}{h^2} & \ldots & 0 \\
\vdots & \ddots & \ddots & \ddots & \ddots \\
0 & \ldots & -\frac{\theta}{h^2} & \frac{2\theta}{h^2} & 1 \\
0 & \ldots & 0 & -\frac{1}{h} & \frac{1}{h}
\end{bmatrix},
\] (4.19)

and
\[
X_2 = \begin{bmatrix}
0 & 0 & 0 & \ldots & 0 \\
\frac{1-\theta}{h^2} & \frac{1}{\tau} - \frac{2(1-\theta)}{h^2} & \frac{1-\theta}{h^2} & \ldots & 0 \\
\vdots & \ddots & \ddots & \ddots & \ddots \\
0 & \ldots & \frac{\theta}{h^2} & \frac{1}{\tau} - \frac{2(1-\theta)}{h^2} & \frac{1}{\tau} - \frac{2(1-\theta)}{h^2} \\
0 & \ldots & 0 & 0 & 0
\end{bmatrix},
\] (4.20)

moreover,
where $\tau$ and $h$ denote the time and the spatial step sizes, respectively.

For the sake of simplicity let us consider linear kinetics, i.e., let $g(u) = u$, and let us denote $\beta = \frac{2}{h^2} + \nu^2$. Then with the second order approximation on the boundary, equation (4.18) can be rewritten as follows (for the Robin type boundary condition see Section 2.5.2).

\[ \tilde{X}_1 u^{\ell+1} = \tilde{X}_2 u^\ell + \tilde{f}, \] (4.22)
\[ \bar{f} = \begin{bmatrix} \frac{FL}{\kappa_{\text{eff}} RT} I(\delta) \\ \vdots \\ 0 \\ \vdots \\ \frac{FL}{\sigma_{\text{eff}} RT} I(\delta) \end{bmatrix}. \] (4.25)

Based on the results of the previous chapters it is easy to see, that the sufficient conditions for the non-negativity of solution at the new time level \((\ell + 1)\) are the following:

\[ \dot{X}^{-1}_1 \geq 0 \text{ and } \dot{X}_2 u^\ell + \bar{f} \geq 0. \] (4.26)

Now considering the structure of the matrix \(\dot{X}_1\) (it is tridiagonal and SDD, i.e., M-matrix, therefore monotone) and the sign of the boundary conditions, we can guarantee the non-negativity of the solution (assuming \(u^\ell \geq 0\)), by the condition \(\dot{X}_2 \geq 0\).

**Theorem 4.4.1** Let us assume that the condition

\[ \frac{1}{(1 - \theta)\beta} \geq \tau \] (4.27)

holds. Then for the one-dimensional problem \((4.22)\), with arbitrary non-negative initial condition, the finite difference method results in a non-negative solution on any time level.

**Theorem 4.4.2** We assume that condition (4.27) holds, then the following estimation is valid on any time level:

\[ 0 \leq u^{\ell+1} \leq \max \left\{ \max u^\ell, H(I(t)) \right\}, \] (4.28)

where \(H(I(t))\) is a function which represents the time-dependent load current density, i.e., the boundary condition.

**Proof:** Let us use the notation of Theorem 2.4.1, i.e., let \(n = \frac{1}{h} + 1\), \(A = \dot{X}_1\), \(F = \dot{X}_2 u^\ell + \bar{f}\) and \(u = u^{\ell+1}\) for any time level \(\ell\). The lower estimate is obvious since the conditions which guarantee the non-negativity of the solution hold. Since, \(\dot{X}_1\) is SDD and monotone, the statements of Theorem 2.4.1 are valid, i.e.,

\[ 0 \leq u^{\ell+1} \leq \max \left\{ 0, \max_{j=1,...,n} \frac{\dot{X}_2 u^\ell + \bar{f}}{\alpha_j(\dot{X}_1)} \right\}. \] (4.29)
Now considering the entries of the matrices and \( \tilde{f} \) we have
\[
\max \left\{ 0, \max_{j=1,\ldots,n} \left\{ \frac{\tilde{X}_2 u^\ell + \tilde{f}}{\alpha_j(\tilde{X}_1)} \right\} \right\} = \max \left\{ 0, \max_{j=2,\ldots,n-1} \left\{ \frac{\tilde{X}_2 u^\ell}{\alpha_j(\tilde{X}_1)}, \max \frac{2\tilde{f}}{h\nu^2} \right\} \right\}.
\tag{4.30}
\]
It is easy to obtain
\[
\|\tilde{X}_2\|_{\infty} = \frac{1}{\tau} - \nu^2(1 - \theta)
\tag{4.31}
\]
and
\[
\alpha_j(\tilde{X}_1) = \frac{1}{\tau} + \theta\nu^2, \quad j = 2, \ldots, n - 1.
\tag{4.32}
\]
Then on the base of Remark 3.5.2 we obtained the following:
\[
0 \leq u^{\ell+1} \leq \max \left\{ \max u^\ell, \max \frac{2\tilde{f}}{h\nu^2} \right\},
\tag{4.33}
\]
where \( \max \frac{2\tilde{f}}{h\nu^2} \) (besides some constant material and numerical parameter) depends only on the load current density, hence, the theorem is proven.

\[\blacksquare\]

\textbf{Remark 4.4.1} For problem (4.18), the theorem above is not valid, since \( X_1 \) is not SDD.

Let us apply to the same simplified problem the linear finite element method for the spatial and \( \theta \)-method for the time discretization of this problem, then the equation
\[
X_1 u^{\ell+1} = X_2 u^\ell + G, \quad \ell = 0, \ldots
\tag{4.34}
\]
presents the fully discretized problem on the given level, where using the notations of Chapter 3, \( X_1 = \frac{1}{\tau} M + \theta(Q + \nu^2 M) \) and \( X_2 = \frac{1}{\tau} M - (1 - \theta)(Q + \nu^2 M) \) and \( G \) represents the boundary conditions.

Similarly for the heat conduction equation in Chapter 3 (3.56 and 3.57), it is easy to get the following theorem.

\textbf{Theorem 4.4.3} Let us assume that the conditions
\[
\frac{1}{\theta \left( \frac{6}{h^2} - \nu^2 \right)} > \tau,
\tag{4.35}
\]
and
\[
\frac{1}{(1 - \theta) \left( \frac{3}{h^2} - \nu^2 \right)} \leq \tau
\tag{4.36}
\]

hold. Then for the one-dimensional problem (4.34), with arbitrary non-negative initial condition, the applied linear finite element method results in a non-negative solution on any time level.

For the solution of the terms in (4.1) see the results of Section 4.6.1.

4.5 Application of Operator Splitting

The operator splitting method is well known and widely used in many fields of modeling [7], [21], [109] for solving time-dependent complex physical problems, where the operators in the partial differential equations describe different sub-processes. In the course of the operator splitting these sub-processes are handled separately, i.e., the original problem can be approached by solving some simpler problems instead of solving one complex problem. This approach was also successfully applied to some other kind of physical problems (air pollution modeling, advection-diffusion problem, etc.) and to a fuel cell model, too, in our work [66].

In this section, we are presenting the general mathematical model of the operator splitting methods, and later we are going to extend the results of the previous section to these methods, too. The detailed analysis of operator splitting methods can be found in [27].

The general abstract mathematical model can be formulated as the Cauchy problem

\[
\frac{dw(t)}{dt} = \sum_{i=0}^{q} A_i w(t), \quad t \in (0, t_{\text{max}}),
\]

\[
w(0) = w_0.
\]  

In the above formulation \( w : \mathbb{R} \to X \) is the unknown (X-valued) function, where \( X \) denotes the space of the possible states (assumed to be a Banach space), and \( w_0 \in X \) is the given initial state (initial condition). Operators \( A_i : X \to X \) are assumed to be known, and they define \( q \) different sub-processes.

Considering a given problem, two questions arise immediately. The first is, how to define the different sub-operators \( A_i \). Spatially? According to the different time-scales? Split between the linear and non-linear sub-processes? The second question is, how to solve the sub-problems. Which numerical methods are able to give the best performance and precision for each sub-process?

As follows, the separation by linear and non-linear processes in problem (4.13) is applied,
Figure 4.3: The first two time steps in the algorithm of the sequential splitting method for both possible cases $SEQ_{12}$ and $SEQ_{21}$

which is quite a natural choice in our case. Then the two operators are the following:

$$P_1u = \partial_\xi \xi u \quad \text{and} \quad P_2u = -\nu^2 g(u), \quad (4.39)$$

where $P_1$ and $P_2$ denotes the so-called diffusion and source operator, respectively. According to (4.37) these operators define two subproblems ($q = 2$), where the sub-problem defined by $P_1$ needs initial and boundary conditions as well, while the problem defined by $P_2$ needs only initial condition.

**Remark 4.5.1** *It is clear that the diffusion operator is linear, while the source operator can be both, linear and non-linear, depending on the applied kinetics.*

Many different types of the splitting method exist, however, in the course of the analysis and numerical experiments sequential splitting and symmetrically weighted splitting methods are applied to problem (4.13).

The procedure of sequential splitting (when we solve the sub-problem of $P_1$ first) is as follows (cf. Fig. 4.3 - $SEQ_{12}$):

1. With the given homogeneous initial condition solve the problem defined by $P_1$ with an arbitrary time step $\tau$.

2. Use the solution of the previous step as an initial condition to solve the problem defined by $P_2$ with the time step $\tau_2 := \frac{\tau}{n_\tau}$, where $n_\tau \in \mathbb{N}$ denotes the subdivision of the time step of the first problem.
3. The solution on the first time level of the splitted problem is the solution of the second step, denoted by $u(\tau, \xi)$.

4. Repeat from the step 1 using the results of step 3 as initial condition.

Let us denote by $\overline{u}(\delta, \xi)$ the solution of the reverse order sequential splitting (see $SEQ_{21}$ from Fig. 4.3). It can be shown that the sequential splitting has the first order of accuracy [27], and it is not symmetrical with respect to the order of the operators, i.e., $u(\delta, \xi) \neq \overline{u}(\delta, \xi)$. The idea of the symmetrically weighted method is to calculate the solutions of the sequential splitting in both orderings, and take their average value as the solution at every time level, i.e.,

$$
\overline{u}(\delta, \xi) = \frac{1}{2}(u_2(\delta, \xi) + u_1(\delta, \xi)),
$$

where $u_2(\delta, \xi)$ and $u_1(\delta, \xi)$ denote the solutions of the sequential splitting methods in both orderings, however, using $\overline{u}(\delta - \tau, \xi)$ as initial condition at the given time level. This method has second order accuracy [27] and it is symmetrical, since the result does not depend on the order of the calculations (see $SYM$ from Fig. 4.4).

Let us solve the problem $P_1$ by using the finite difference $\theta$-method (as in the previous section), where the discretization matrices are exactly the same as in (4.19) and (4.20), only the right-hand side is different, where the source term disappears. We solve $P_2$ explicitly (let $\tau_2 = \tau$), i.e.,

$$
u_{\ell+1} = u_{\ell} - \tau v^2 g(u_{\ell}),
$$

(4.41)
where \( u^\ell \) now denotes the initial condition of this problem, which is the solution of the previous step of the given splitting method. Then the following theorem is valid.

**Theorem 4.5.1** Assume that the conditions

\[
\frac{h^2}{2(1-\theta)} \geq \tau \quad \text{and} \quad \frac{u(\delta, \xi)}{\nu^2 g(u(\delta, \xi))} \geq \tau_2
\]  

(4.42)

hold for all \( u(\delta, \xi) \neq 0 \), where \( (\delta, \xi) \in (0, \delta_{\text{max}}) \times [0, 1] \). Then for the one-dimensional problem (4.13), with an arbitrary non-negative initial condition, the presented splitting methods, using \( \theta \)-method for the problem of \( P_1 \) and the explicit method (4.41) for \( P_2 \), results in a non-negative solution at any time level. (For \( u(\delta, \xi) = 0 \) the second condition is considered in the sense of limits.)

**Proof:** To prove the non-negativity of the numerical solution obtained by the splitting methods it is sufficient to show the non-negativity of the solution of the problem of \( P_1 \) and \( P_2 \) under the given assumptions with an arbitrary non-negative initial condition. The first condition follows from the structure of matrix (4.20), i.e., we can guarantee the non-negativity of the solution of the problem of \( P_1 \) by the condition \( X_2 \geq 0 \). It is easy to check if the second condition holds, the left-hand side of (4.41) is non-negative, hence, the theorem is proven. \( \blacksquare \)

**Remark 4.5.2** According to Theorem 4.5.1 and the structure of \( X_2 \) in (4.20), if \( u(\delta, \xi) \equiv 0 \) on a given time level, then there is no condition on the time step-size for the next step, i.e., using homogeneous initial condition the method above results in a non-negative solution in the first time step for any time step-size.

**Remark 4.5.3** Obtaining a similar theorem for the method when \( P_1 \) is solved by finite element \( \theta \)-method is straightforward.

**Remark 4.5.4** Let us solve the problem of \( P_2 \) by the following implicit method:

\[
u^{\ell+1} + \tau \nu^2 g(u^{\ell+1}) = u^\ell.
\]  

(4.43)

Then, considering the sign and the monotonicity of \( g(u) \), in the previous theorem only the first condition is required to preserve the non-negativity of the solutions obtained by the presented splitting methods. (The right-hand side of (4.43) is a positive constant and the left-hand side is a monotone function of the unknown, i.e., the solution is positive and unique.)
4.6 Heterogeneous Model

In this section we will obtain an explicit equation for the overpotential \( \eta(t, x) \) by eliminating the term \( \phi_2(t, x) \) in (4.8) without assuming constant material and kinetic coefficients. This generalizes the result in the previous section, where this has been done in case of constant coefficients \( \kappa_{\text{eff}} \) and \( \sigma_{\text{eff}} \).

In the major part of the following derivation, for simplicity, we skip the variables \( t \) and \( x \). Using (4.6) and taking the derivative of (4.7) we obtain that

\[
\partial_x(\sigma_{\text{eff}} \partial_x \phi_1) = \partial_x i_1 = \partial_x i_2 = -\partial_x (\kappa_{\text{eff}} \partial_x \phi_2), \tag{4.44}
\]

which, together with the definition (4.4) of \( \eta(t, x) \) gives

\[
\partial_x(\sigma_{\text{eff}} \partial_x \phi_2 + \kappa_{\text{eff}} \partial_x \phi_2) = \partial_x(\sigma_{\text{eff}} \partial_x \phi_2) - \partial_x(\sigma_{\text{eff}} \partial_x \phi_1) = -\partial_x(\sigma_{\text{eff}} \partial_x \eta). \tag{4.45}
\]

Since the two derivatives in (4.45) are equal, we obtain

\[
(\kappa_{\text{eff}}(x) + \sigma_{\text{eff}}(x)) \partial_x \phi_2(t, x) = -\sigma_{\text{eff}}(x) \partial_x \eta(t, x) + (\kappa_{\text{eff}}(0) + \sigma_{\text{eff}}(0)) \partial_x \phi_2(t, 0) + \sigma_{\text{eff}}(0) \partial_x \eta(t, 0) \tag{4.46}
\]

where in the second line the boundary conditions (4.12) have been used twice. Using (4.45) and (4.46), we rewrite the left-hand side of equation (4.8) as

\[
\partial_x(\kappa_{\text{eff}} \partial_x \phi_2) = \partial_x \left( \frac{\kappa_{\text{eff}}}{\kappa_{\text{eff}} + \sigma_{\text{eff}}} \kappa_{\text{eff}} \partial_x \phi_2 + \frac{\kappa_{\text{eff}}}{\kappa_{\text{eff}} + \sigma_{\text{eff}}} \sigma_{\text{eff}} \partial_x \phi_2 \right)
\]

\[
= \partial_x \left( \frac{\kappa_{\text{eff}}}{\kappa_{\text{eff}} + \sigma_{\text{eff}}} \right) (\kappa_{\text{eff}} \partial_x \phi_2 + \sigma_{\text{eff}} \partial_x \phi_2)
\]

\[
+ \frac{\kappa_{\text{eff}}}{\kappa_{\text{eff}} + \sigma_{\text{eff}}} \partial_x (\kappa_{\text{eff}} \partial_x \phi_2 + \sigma_{\text{eff}} \partial_x \phi_2)
\]

\[
= -\partial_x \left( \frac{\kappa_{\text{eff}}}{\kappa_{\text{eff}} + \sigma_{\text{eff}}} \right) (\sigma_{\text{eff}} \partial_x \eta - I(t)) - \frac{\kappa_{\text{eff}}}{\kappa_{\text{eff}} + \sigma_{\text{eff}}} \partial_x (\sigma_{\text{eff}} \partial_x \eta). \tag{4.47}
\]

Substituting (4.47) into the left-hand side of (4.8), it becomes the explicit equation

\[
a C_{\text{dl}} \partial_t \eta(t, x) = \partial_x \left( \frac{\kappa_{\text{eff}}}{\kappa_{\text{eff}} + \sigma_{\text{eff}}} \right) (-I(t) + \sigma_{\text{eff}} \partial_x \eta(t, x))
\]

\[
+ \frac{\kappa_{\text{eff}}}{\kappa_{\text{eff}} + \sigma_{\text{eff}}} \partial_x (\sigma_{\text{eff}} \partial_x \eta(t, x)) - a_i \beta \left( \alpha \frac{F}{RT} \eta(t, x) \right). \tag{4.48}
\]
for the unknown \( \eta(t, x) \), where also the functions \( a, i_0, C_{\text{dl}}, \kappa_{\text{eff}} \) and \( \sigma_{\text{eff}} \) depend on \( (x) \) with \( t \in (0, t_{\text{max}}) \) and \( x \in (0, L) \). For the corresponding initial-boundary value problem we use the initial value

\[
\eta(0, x) = 0, \quad x \in (0, L),
\]

and \((4.48)\) is equipped with the Neumann type boundary conditions in \((4.12)\).

Based on \((4.46)\), we can express \( \phi_2(t, x) \) as

\[
\partial_x \phi_2(t, x) = \frac{1}{\kappa_{\text{eff}}(x) + \sigma_{\text{eff}}(x)} \left( I(t) - \sigma_{\text{eff}}(x) \partial_x \eta(t, x) \right),
\]

and consequently, by the assumption \( \phi_2(t, 0) = 0 \) (see the explanation after \((4.4)\) we have

\[
\phi_2(t, x) = \int_0^x \left( -\frac{\sigma_{\text{eff}}(t, s)}{\kappa_{\text{eff}}(t, s) + \sigma_{\text{eff}}(t, s)} \partial_s \eta(t, s) + \frac{1}{\kappa_{\text{eff}}(t, s) + \sigma_{\text{eff}}(t, s)} I(t) \right) \, ds.
\]

Therefore, according to \((4.5)\) we can give the potential loss \( V^* \) at the anode as

\[
V^*(t) = \eta(t, L) + \phi_2(t, L)
= \eta(t, L) + \int_0^L \left( -\frac{\sigma_{\text{eff}}(t, s)}{\kappa_{\text{eff}}(t, s) + \sigma_{\text{eff}}(t, s)} \partial_s \eta(t, s) + \frac{1}{\kappa_{\text{eff}}(t, s) + \sigma_{\text{eff}}(t, s)} I(t) \right) \, ds.
\]

This completes the computation of the right-hand side of \((4.1)\), and the desired quantity \( E_{\text{cell}}(t) \) can be given.

**Remark 4.6.1** For the sake of simplicity, we did not apply the time dependence of the material and kinetic parameters in the course of the analysis, however, it is easy to extend it to this case, too.

### 4.6.1 Numerical Solution and its Properties

On the next pages we are discussing in detail the numerical solution of the terms in \((4.1)\). The most involved step is the computation of the overpotential \( \eta(t, x) \).

To solve \((4.2)\) numerically, we apply the well-known Newton-Raphson method (assuming that \( f_a \in C^1(\mathbb{R}) \)). The approximation of \( \eta^a(t) \) is the root of the function \( f_a \), which is defined by

\[
f_a(s) = i_0^a(t) \left[ \exp \left( \alpha_s^a \frac{F}{RT} s \right) - \exp \left( -\alpha_s^a \frac{F}{RT} s \right) \right] - I(t),
\]

with the derivative

\[
f'_a(s) = \frac{i_0^a(t) F}{RT} \left[ \alpha_s^a \exp \left( \alpha_s^a \frac{F}{RT} s \right) + \alpha_s^a \exp \left( -\alpha_s^a \frac{F}{RT} s \right) \right] \geq 0.
\]
Let \( \eta_j^a(t) \) denote the \( j^{th} \) iteration for \( \eta^a \) initiated from \( s = \eta_0^a \). To obtain an initial guess, one can observe (using also the demand that \( \eta^a \) is positive) that the second term on the right-hand side of (4.2) is near to zero. Omitting this term, one can easily express \( \eta_0^a \) as

\[
\eta_0^a = \frac{a_a F}{RT} \ln \left( \frac{I(t)}{i_0^a} \right). \tag{4.55}
\]

At each iteration step \( j = 0, 1, \ldots \), we define \( \eta_{j+1}^a \) as

\[
\eta_{j+1}^a = \eta_j^a - \frac{f_a(\eta_j^a)}{f_a'(\eta_j^a)}. \tag{4.56}
\]

The iteration step above is repeated until

\[
|\eta_{j+1}^a - \eta_j^a| < \varepsilon, \tag{4.57}
\]

where \( \varepsilon \) denotes the requested tolerance or precision.

For the numerical solution of (4.48) we rewrite it by introducing the function \( S(x) \) and the constant \( K \) with

\[
S(x) = \frac{\kappa_{\text{eff}}(x)}{\kappa_{\text{eff}}(x) + \sigma_{\text{eff}}(x)} \quad \text{and} \quad K = \frac{RT}{\alpha F}.
\]

Let us introduce a new unknown function \( u(t, x) = \frac{\eta(t, x)}{K} \). Then one can rewrite (4.48) as

\[
\begin{align*}
\partial_t u(t, x) &= \partial_x(S(t, x)) \frac{1}{a C_\text{dl}(x)} \left( -\frac{I(t)}{K} + \sigma_{\text{eff}} \partial_x u(t, x) \right) \\
&\quad + \frac{S(t, x)}{a C_\text{dl}(x)} \partial_x(\sigma_{\text{eff}} \partial_x u(t, x)) - \frac{i_0}{C_\text{dl}(x)K} g(u(t, x)).
\end{align*} \tag{4.58}
\]

We solve the corresponding initial-boundary value problem applying an implicit-explicit Euler method. The method is explicit in the source term, therefore iterations for solving nonlinear problems can be avoided. At the same time, the spatial derivative on the right-hand side of (4.58) is approximated using an implicit scheme, which maintains the stability of the time stepping. This approach provides a good balance between accuracy and relatively low computational costs.

For the discretization of the interval \((0, L)\) we use an equidistant grid of size \( h = \frac{L}{n} \) and the time step is denoted by \( \tau \). We use the notation \( u_{j}^{\ell} \) for the approximation of \( u(\ell \tau, jh) \). The same convention is used for the other functions: the upper index \( \ell \) and the lower index \( j \) yields the approximation at time \( \ell \tau \) and at the spatial coordinate \( jh \), respectively.
In concrete terms, we use the following scheme to discretize (4.58):

\[
\frac{u_j^{\ell+1} - u_j^{\ell}}{\tau} = \frac{1}{a_j C_{dj}} \left( \frac{S_j - S_{j-1}}{h} \left( -I_1^{\ell+1} \frac{1}{K} + \frac{u_j^{\ell+1} - u_{j-1}^{\ell+1}}{h} \right) \right)
+ \frac{S_j}{a_j C_{dj}} \left( \frac{\sigma_{j+\frac{1}{2}}}{h} \left( \frac{u_{j+1}^{\ell+1} - u_j^{\ell+1}}{h} \right) - \frac{\sigma_{j-\frac{1}{2}}}{h} \left( \frac{u_j^{\ell+1} - u_{j-1}^{\ell+1}}{h} \right) \right) - \frac{i_{0j}}{C_{dj} K} g(u_j^{\ell})
\]

(4.59)

where we have used the notation

\[
\sigma_{j+\frac{1}{2}} = \frac{\sigma_{j+1} + \sigma_j}{2}
\]

(4.60)

Based on the first order approximations

\[
\frac{h I_1^{\ell+1}}{\kappa_{d0} K} = u_0^{\ell+1} - u_1^{\ell+1}
\]

(4.61)

and

\[
\frac{h I_1^{\ell+1}}{\sigma_{d0} K} = u_n^{\ell+1} - u_{n-1}^{\ell+1}
\]

(4.62)

for \(\ell = 1, 2, \ldots\), the discretization of problem (4.12)–(4.48) becomes the following system of algebraic equations:

\[
X u^{\ell+1} = (X_1 - X_2) u^{\ell+1} = f(u^{\ell}),
\]

(4.63)

where

\[
f(u^{\ell}) = \begin{bmatrix}
\frac{h I_1^{\ell+1}}{\kappa_{d0} K} \\
\vdots \\
\frac{h I_1^{\ell+1}}{\sigma_{d0} K} \\
\frac{h I_1^{\ell+1}}{\sigma_{d1} K} \\
\vdots \\
\frac{h I_1^{\ell+1}}{\sigma_{dn} K}
\end{bmatrix}
\]

\[
X_1 = \begin{bmatrix}
1 & -1 & 0 & \ldots & \ldots & \ldots & 0 \\
\vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\
0 & \ldots & -\sigma_{d1-\frac{1}{2}} b_j & 1 + s_j b_j & -\sigma_{d1+\frac{1}{2}} b_j & \ldots & 0 \\
\vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\
0 & \ldots & \ldots & \ldots & \ldots & 0 & -1 & 1
\end{bmatrix}
\]

92
\[ s_j = \sigma_{\text{eff}, j-\frac{1}{2}} + \sigma_{\text{eff}, j+\frac{1}{2}}, \quad (4.66) \]

\[ b_j = \frac{\tau S_j}{a_j C_{\text{dl}, j} h^2}, \quad (4.67) \]

\[
X_2 = \begin{bmatrix}
0 & 0 & 0 & \ldots & \ldots & \ldots & 0 \\
-d_2 & d_2 & 0 & \ldots & \ldots & \ldots & 0 \\
0 & -d_3 & d_3 & \ldots & \ldots & \ldots & 0 \\
\vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\
\vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\
0 & \ldots & \ldots & \ldots & 0 & 0 & 0 \\
\end{bmatrix}, \quad (4.68)
\]

and

\[ d_j = \frac{\sigma_{\text{eff}, j} \tau (S_j - S_{j-1})}{a_j C_{\text{dl}, j} h^2}. \quad (4.69) \]

Now we use the assumption that the solution of \((4.12) - (4.48)\) is sufficiently smooth such that the forthcoming Taylor expansions are justified.

To verify the consistency of the scheme in \((4.59)\), we first use the Taylor expansion of the left-hand side in \(t\) at \((t_{\ell + 1}, x_j)\), which gives

\[ \frac{u_j^{\ell + 1} - u_j^\ell}{\tau} = \partial_t u(t_{\ell + 1}, x_j) - \frac{1}{2} \tau \partial_{tt} u(t_{\ell + 1}, x_j) + O(\tau^2). \quad (4.70) \]

Similarly, the Taylor expansions in \(x\) at \((t_{\ell + 1}, x_j)\) imply the identities

\[ \frac{S_j - S_{j-1}}{h} = \partial_x S(x_j) - \frac{1}{2} h \partial_{xx} S(x_j) + O(h^2), \quad (4.71) \]

and

\[ \frac{u_j^{\ell + 1} - u_{j-1}^\ell}{h} = \partial_x u(t_{\ell + 1}, x_j) - \frac{1}{2} h \partial_{xx} u(t_{\ell + 1}, x_j) + O(h^2), \quad (4.72) \]
Using Taylor expansions in $x$ which can be used to rewrite the second term on the right-hand side of (4.59) such that the first term of the right-hand side of (4.59) becomes

\[
\frac{1}{a_j C_{dij}} \frac{S_j - S_{j-1}}{h} \left( -I^{\ell+1} \frac{1}{K} + \sigma_{eff, j} \frac{u^{\ell+1}_j - u^{\ell+1}_{j-1}}{h} \right)
\]

\[
= \frac{1}{a_j C_{dij}} (\partial_x S(x_j) + O(h)) \left( -I^{\ell+1} \frac{1}{K} + \sigma_{eff}(x_j) \partial_x u(t_{\ell+1}, x_j) + O(h) \right)
\]

\[
= \frac{1}{a_j C_{dij}} \partial_x S(x_j) \left( -I^{\ell+1} \frac{1}{K} + \sigma_{eff}(x_j) \partial_x u(t_{\ell+1}, x_j) \right) + O(h).
\]

Using Taylor expansions in $x$ at $(t_{\ell+1}, x_j)$ we also obtain

\[
\sigma_{eff, j + \frac{1}{2}} = \sigma_{eff}(x_j) + \frac{1}{2} h \partial_x \sigma_{eff}(x_j) + \frac{1}{4} h \partial_{xx} \sigma_{eff}(x_j) + O(h^3),
\]

\[\text{(4.74)}\]

\[
\sigma_{eff, j - \frac{1}{2}} = \sigma_{eff}(x_j) - \frac{1}{2} h \partial_x \sigma_{eff}(x_j) + \frac{1}{4} h \partial_{xx} \sigma_{eff}(x_j) + O(h^3),
\]

\[\text{(4.75)}\]

\[
\frac{u^{\ell+1}_j - u^{\ell+1}_{j-1}}{h} = \partial_x u(t_{\ell+1}, x_j) + \frac{1}{2} h \partial_{xx} u(t_{\ell+1}, x_j) + \frac{1}{6} h^2 \partial_{xxx} u(t_{\ell+1}, x_j) + O(h^3),
\]

\[\text{(4.76)}\]

\[
\frac{u^{\ell+1}_j - u^{\ell+1}_{j-1}}{h} = \partial_x u(t_{\ell+1}, x_j) - \frac{1}{2} h \partial_{xx} u(t_{\ell+1}, x_j) + \frac{1}{6} h^2 \partial_{xxx} u(t_{\ell+1}, x_j) + O(h^3),
\]

\[\text{(4.77)}\]

which can be used to rewrite the second term on the right-hand side of (4.59) as

\[
\frac{S_j}{a_j C_{dij}} \left( \frac{\sigma_{eff, j + \frac{1}{2}}}{h} \left[ \frac{u^{\ell+1}_j - u^{\ell+1}_{j-1}}{h} \right] - \frac{\sigma_{eff, j - \frac{1}{2}}}{h} \left[ \frac{u^{\ell+1}_j - u^{\ell+1}_{j-1}}{h} \right] \right)
\]

\[\text{(4.78)}\]

\[
= \frac{S_j}{a_j C_{dij}} \sigma_{eff}(x_j) \partial_{xx} u(t_{\ell+1}, x_j) + \partial_x \sigma_{eff}(x_j) \partial_x u(t_{\ell+1}, x_j) + O(h^2)
\]

\[
= \frac{S_j}{a_j C_{dij}} \partial_x (\sigma_{eff}(x_j) \partial_x u(t_{\ell+1}, x_j)) + O(h^2).
\]

In the same way, in the last term of the right-hand side of (4.59) for $g(u^\ell)$:

\[
g \left( u^\ell_j \right) = g \left( u(t_{\ell+1}, x_j) \right) - \tau \partial_t u(t_{\ell+1}, x_j) g'(u(t_{\ell+1}, x_j)) + O(\tau^2)
\]

\[= g \left( u(t_{\ell+1}, x_j) \right) + O(\tau).
\]

\[\text{(4.79)}\]

Taking the sum of (4.73), (4.78) and (4.79) we obtain the following expansion for the
right-hand side of \((4.59)\):

\[
\frac{1}{a_j C_{dl,j}} \partial_x S(x_j) \left( -\frac{I(t+1)}{K} + \sigma_{\text{eff}}(x_j) \partial_x u(t_{t+1}, x_j) \right) \\
+ \frac{S(x_j)}{a_j C_{dl,j}} \partial_x \left( \sigma_{\text{eff}}(x_j) \partial_x u(t_{t+1}, x_j) \right) \\
+ \frac{S(x_n)}{a_j C_{dl,j}} g(u(t_{t+1}, x_j)) + O(h) + O(h^2) + O(\tau),
\]

which compared with the Taylor expansion \((4.72)\) shows that the approximation error of the scheme in \((4.59)\) is of order \(O(\tau + h)\). This proves the consistency of the presented scheme \((4.59)\).

It is not possible to guarantee the non-negativity preservation of the method for the linear case by the same way, which was used for the macro-homogeneous case, however, it is easy to obtain the following theorem.

**Theorem 4.6.1** We assume that the following conditions hold:

\[
h \frac{a(x)C_{dl}(x)}{\sigma_{\text{eff}}(x)} \partial_x S(x) \geq \tau, \tag{4.81}
\]

if \(\partial_x S(x) \neq 0\) and

\[
\frac{KC_{dl}(x)u(t, x)}{I(t) a(x) \partial_x S(x) + i_0(x) g(u(t, x))} \geq \tau. \tag{4.82}
\]

for all \((t, x) \in (0, t_{\max}) \times [0, L]\), where the denominator is considered to be non-zero, i.e., \(-\partial_x S(x) \neq I(t)a(x)i_0(x)g(u(t, x))\). For \(u(t, x) = 0\) the second condition in \((4.81)\) is considered in the sense of limits. Then \(u^t \geq 0\) yields \(f(u^t) \geq 0\), and \((4.81)\) results in a non-negative solution \(u^{t+1}\) with any non-negative initial condition \(u^t\). If \(\partial_x S(x) = 0\), then only the second condition is needed for the non-negativity preservation of the method.

**Proof:** The second condition \((4.82)\) guarantees the non-negativity of the right-hand side (considering implicit discretization for \(I(t)\) and the structure of \((4.64)\), furthermore, it is easy to see that if the first condition \((4.81)\) is fulfilled, then

\[
d_j \leq \frac{1}{2} \quad \text{for all} \ j = 2, \ldots, n - 1, \tag{4.83}
\]

which implies that \(X\) is SDD and tridiagonal, i.e, an M-matrix, therefore, it is monotone, hence the statements of the theorem are proven.

\[
95
\]
Remark 4.6.2 According to Theorem 4.6.1 if \(-\partial_x S(x) = I(t) a(x) i_0(x) g(u(t, x))\) for all \(x \in [0, L]\) at a given time level, then there is no condition on the time step size for the next step.

Remark 4.6.3 If the load current is zero and/or the conductivities are spatially constants, then instead of (4.81), the condition
\[
\frac{KC_{\delta\delta}(x) u(t, x)}{i_0(x) g(u(t, x))} \geq \tau
\]
holds for all \(u(t, x) \neq 0\), where \((t, x) \in (0, t_{\text{max}}) \times [0, L]\) guarantees the non-negativity of the solution of problem (4.63), and if \(u(\delta, \xi) \equiv 0\) on a given time level, then there is no condition on the time step size for the next step. For \(u(t, x) = 0\) the condition (4.84) is considered in the sense of limits.

Remark 4.6.4 It is worth emphasizing that the non-negativity of the solution cannot be guaranteed in every case by the help of the previous theorem. As an example let us apply a monotonously decreasing function as the conductivity at the solution phase \((\kappa_{\text{eff}}(x))\), which implies \((\sigma_{\text{eff}}\text{ is considered to be constant})\) negative \(\partial_x S(x)\). Then, with homogeneous initial conditions the left-hand side of (4.82) is negative for any non-zero load current density \((I(t))\), therefore, there is no such time step size \(\tau > 0\) which can guarantee the non-negativity preservation of the method.

Remark 4.6.5 Using the implicit backward discretization (explicit in the source term) instead of the \(\theta\)-method presented in (4.18) for the macro-homogeneous case, we can obtain the following condition to preserve the non-negativity of the solution:
\[
\frac{u(\delta, \xi)}{\nu^2 g(u(\delta, \xi))} \geq \tau, \quad (4.85)
\]
for all \(u(\delta, \xi) \neq 0\), where \((\delta, \xi) \in (0, \delta_{\text{max}}) \times [0, 1]\), which is equivalent to the second expression of (4.42). If \(u(\delta, \xi) = 0\), then (4.85) must be valid in the sense of limits.

4.6.2 Test Problem

In this part we test the accuracy of the presented numerical method (4.63). We have analyzed a non-trivial model problem with spatially heterogeneous conductivity parameters. Based on real measurements we have the approximate values \(\kappa_{\text{eff}} \approx 0.002\) and \(\sigma_{\text{eff}} \approx 1.8\) and accordingly, we define
\[
\kappa_{\text{eff}}(t, x) = 0.002 - 0.001 x \quad \text{and} \quad \sigma_{\text{eff}}(t, x) = 1.8 + 0.001 x. \quad (4.86)
\]
Consequently,
\[ \kappa_{\text{eff}} \sigma_{\text{eff}} = 1.802 \quad \text{and} \quad \frac{\kappa_{\text{eff}}}{\sigma_{\text{eff}} + \sigma_{\text{eff}}}(t, x) = \frac{2 - x}{1802}. \quad (4.87) \]

If the analytic solution of the governing equation (4.48) is
\[
\eta(t, x) = \frac{1}{4} t \left(1 + \left(x - \frac{1.801}{1.803}\right)^2\right) \Leftrightarrow \partial_x \eta(t, x) = \frac{t}{2} \left(x - \frac{1.801}{1.803}\right), \quad (4.88)
\]
we can verify that the equalities
\[
\partial_x \eta(t, 0) = -\frac{1}{\kappa_{\text{eff}}(t, 0)} I(t) \quad \text{and} \quad \partial_x \eta(t, L) = \frac{1}{\sigma_{\text{eff}}(t, 1)} I(t) \quad (4.89)
\]
hold true, where \( I(t) = \frac{1.801}{1.803} \cdot 10^{-3} t \). These show that the boundary conditions in (4.12) are satisfied. Inserting (4.86), (4.88) and (4.89) into (4.48), a straightforward calculation leads to
\[
C_{\text{dl}}(x) = \frac{4t}{a \cdot 3604 \cdot \left(1 + \left(x - \frac{1.801}{1.803}\right)^2\right)} \left(-0.002x^2 - 3.595x + 5.398\right) - \frac{4i_0}{1 + \left(x - \frac{1.801}{1.803}\right)^2} g \left(\frac{F}{RT} \cdot \frac{t}{4} \left(1 + \left(x - \frac{1.801}{1.803}\right)^2\right)\right). \quad (4.90)
\]

Using the above function \( C_{\text{dl}} \) and the constants in Table 4.1 the governing equation (4.48) is given explicitly, which, imposed with boundary conditions (4.89) and initial condition \( \eta(0, x) = 0 \), has the analytic solution given in (4.88). The computation has been performed on 201 nodes over 100 time steps. In Figure 4.5 one can compare the analytic solution with the result of the numerical approximation, and the computational error is depicted in Figure 4.6.
Figure 4.5: Analytic solution (4.88) of (4.48) (continuous line) and the numerical approximation (dashed line) using the method in Section 4.6.1 at $t = 1$ after 100 time steps. The parameters are given in (4.86) and Table 4.1.

Figure 4.6: Error in the computations from Figure 4.5

4.6.3 Coupled Diffusion Problem

In an extended version of the heterogenous model, the diffusion of oxygen (see Fig. 4.1) at the cathode is also considered. In this case, there is a need to solve the following problem coupled to the macro-homogeneous (see Section 4.4) or the heterogeneous problem (see Section 4.6)

$$
\partial_t [O_2](t, x) = \partial_x (D(x) \partial_x [O_2](t, x)) - \frac{i_f(t, x)}{4F},
$$

(4.91)
where for the faradaic current at the cathode, the equation

\[ i_f(t, x) = a_i0g(u(t, x)) \]  \quad (4.92)

holds, where \( u(t, x) \) is the solution of system (4.48).

At \( t = 0 \) the oxygen concentration is equal to the concentration of the source, which yields

\[ [O_2](0, x) = [O_2]_L \]  \quad (4.93)

as initial condition.

The outward transport of the oxygen is assumed to be zero at the common boundary of the cathode and the membrane \((x = 0)\), furthermore, the concentration at the current collector \((x = L)\) is equal to the concentration of the oxygen source, i.e., we have the following boundary conditions:

\[
\begin{align*}
\partial_x [O_2](t, 0) &= 0, \ t \in (0, T), \\
[O_2](t, L) &= [O_2]_L, \ t \in (0, T).
\end{align*}
\]  \quad (4.94)

By applying an implicit, first order finite difference method (explicit in the source term), the numerical solution of the problem above is the solution of the following system of equations (using the notations of Section 4.6.1)

\[ X_c [O_2]^{t+1} = f_c([O_2]^t), \]  \quad (4.95)

where

\[
X_c = \begin{bmatrix}
\frac{1}{\tau} & -\frac{1}{\tau} & 0 & \ldots & \ldots & \ldots & 0 \\
\frac{1}{\tau} & -\frac{1}{\tau} & 0 & \ldots & \ldots & \ldots & 0 \\
\vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\
0 & \ldots & -\frac{\tau}{h^2}D_{j-\frac{1}{2}} & 1 + \frac{\tau}{h^2}D_j & -\frac{\tau}{h^2}D_{j+\frac{1}{2}} & \ldots & 0 \\
\vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\
0 & \ldots & \ldots & \ldots & \ldots & 0 & 0 & 1
\end{bmatrix}
\]  \quad (4.97)
and

\[ \bar{D}_j = D_{j-\frac{1}{2}} + D_{j+\frac{1}{2}}. \]  

(4.98)

We have seen in \[(4.91)\] that the consumption of the oxygen at the cathode depends on the faradaic current, defined in \[(4.92)\]. The opposite direction of the coupling depends on the applied kinetics. In the case of the Butler–Volmer and the linear kinetics the function \(g(u(t, x))\) should be replaced with the following:

\[ \tilde{g}(u(t, x)) = \left( \frac{[O_2](t, x)}{[O_2]_{\text{ref}}} \right)^\gamma g(u(t, x)), \]  

(4.99)

where \(\gamma\) is the reaction order (in the case of oxygen its value is 1). In the case of the diffusion kinetics, only the parameter \(j_D\) changes, and it can be calculated by the formula

\[ j_D(t, x) = 4F[O_2](t, x) \frac{D_2(x)}{L_2(x)}, \]  

(4.100)

where \(D_2\) is the diffusion coefficient of the oxygen in the solution phase and \(L_2\) is the characteristic size of the agglomerates.

**Remark 4.6.6** The non-negativity preservation of the numerical solution of \[(4.91)\] cannot be guaranteed by any condition on the numerical method for the coupled problem \(10c\) since the source term in \[(4.91)\] does not depend on the unknown variable. Hence, in the course of the numerical modeling, the values of the concentration must be kept non-negative by an additional constraint.

**Remark 4.6.7** Let us assume that for the multiplier at the right-hand side of \[(4.99)\] the following inequalities hold:

\[ 0 \leq \left( \frac{[O_2](t, x)}{[O_2]_{\text{ref}}} \right)^\gamma \leq 1. \]  

(4.101)

Then it is easy to see that the results for the non-negativity preservation with linear kinetics in the previous sections hold, however, the diffusion case still needs to be analyzed.

### 4.7 Numerical Experiments

In this section some numerical experiments for the described fuel cell models (see Sections 4.4 and 4.6) are presented. In the course of the experiments we have applied the finite difference \(\theta\)-methods which were presented in Section 4.4 for the macro-homogeneous and implicit-explicit Euler method for the heterogeneous model. As we did in the previous
Chapters, the numerical experiments are denoted by some ID which consists of some uppercase letters and numbers. In the course of the numerical methods, we have applied MATLAB® for the macro-homogeneous models (FCM1-FCM5, PREC, SEQ12, SEQ21, SYM, SYM2-SYM4) and we applied for FCH1-FCH2 our proprietary software, named CellSim which is based on the heterogeneous model and written in Microsoft Visual C++ 2010 Express. The recent version of the application (CellSim 1.23) has the following features:

- it is able to treat heterogeneous parameter distribution step-wisely,
- applies Richardson-extrapolation [26] on demand,
- results are easily processable in Matlab® and in spreadsheet programs,
- possibility of batch running,
- xml based parameter file.

For the macro-homogeneous model we have applied some space and time transformation (4.15) in order to get the canonical form (4.13), however, in what follows the results are presented in real time and space, i.e., in the course of the numerical solution based on (4.15) inverse transformations were applied.

For the numerical experiments FCM1-FCM5 we considered linear kinetics (4.9), and the applied material and numerical parameters can be found in Table 4.2 and 4.3, where according to condition (4.27) we denote

$$\tau^- = \frac{1}{(1 - \theta)\beta},$$  \hspace{1cm} (4.102)

$$\eta^- = \min_{(\ell, j \mathbin{\underline{\eta}})} (\eta(\ell, j \mathbin{\underline{\eta}})), \quad j = 1, \ldots, N \quad \ell = 0, \ldots, \ell_{\text{max}},$$  \hspace{1cm} (4.103)

and

$$\eta^+ = \max_{(\ell, j \mathbin{\overset{\eta}{\Uparrow}})} (\eta(\ell, j \mathbin{\overset{\eta}{\Uparrow}})), \quad j = 1, \ldots, N \quad \ell = 0, \ldots, \ell_{\text{max}},$$  \hspace{1cm} (4.104)

i.e., the minimum and the maximum of the numerical solution on the given domain, moreover $I_4(\ell \mathbin{\underline{\tau}})$ and $I_5(\ell \mathbin{\overset{\tau}{\Uparrow}})$ in discrete time denote the following functions:

$$I_4(\ell \mathbin{\underline{\tau}}) = \begin{cases} 
1A & \text{if } 0 \leq \ell < 30 \\
0.5A & \text{if } 30 \leq \ell < 60 \\
1A & \text{if } 60 \leq \ell < 90 \\
0.5A & \text{if } 90 \leq \ell \leq \ell_{\text{max}} = 120 
\end{cases},$$  \hspace{1cm} (4.105)
PROTON EXCHANGE MEMBRANE FUEL CELLS

<table>
<thead>
<tr>
<th>ID</th>
<th>$\ell_{\text{max}}$</th>
<th>$\tau$</th>
<th>$\tau_-$</th>
<th>$\eta_-$</th>
<th>$\eta_+$</th>
<th>$\eta_{\text{eff}}$</th>
<th>$I(\ell\tau)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCM1</td>
<td>60</td>
<td>1.00E-03</td>
<td>7.07E-03</td>
<td>0</td>
<td>2.72E-02</td>
<td>3.90E-02</td>
<td>1</td>
</tr>
<tr>
<td>FCM2</td>
<td>60</td>
<td>3.00E-02</td>
<td>7.07E-03</td>
<td>-3.29E-04</td>
<td>2.89E-02</td>
<td>3.90E-02</td>
<td>1</td>
</tr>
<tr>
<td>FCM3</td>
<td>60</td>
<td>5.00E-02</td>
<td>2.12E-02</td>
<td>0</td>
<td>2.72E-02</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>FCM4</td>
<td>120</td>
<td>1.00E-02</td>
<td>7.07E-03</td>
<td>0</td>
<td>2.74E-02</td>
<td>3.90E-02</td>
<td>1</td>
</tr>
<tr>
<td>FCM5</td>
<td>120</td>
<td>1.00E-03</td>
<td>7.07E-03</td>
<td>0</td>
<td>5.37E-02</td>
<td>7.89E-02</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 4.2: Time steps of the numerical models, total current density, bounds, negative results, extreme values and estimations

\[
I_5(\ell\tau) = \begin{cases} 
1A & \text{if } 0 \leq \ell < 30 \\
0.5A & \text{if } 30 \leq \ell < 60 \\
2A & \text{if } 60 \leq \ell < 90 \\
0.5A & \text{if } 90 \leq \ell \leq \ell_{\text{max}} = 120 
\end{cases}
\] (4.106)

The last column of Table 4.2 refers to the estimation of the numerical solution in Theorem 4.33, i.e., considering higher conductivity in the solid phase

\[
\eta_+ = \max \left\{ \max_{\ell=0,\ldots,\ell_{\text{max}}} \left( \max_{j=1,\ldots,N} \eta(\ell\tau, jh) \right), \max_{\ell=0,\ldots,\ell_{\text{max}}} \frac{2FL}{h\nu^2\kappa_{\text{eff}}RT} I(\ell\tau) \right\} .
\] (4.107)

It is worth emphasizing that the estimation can change in time, i.e., for experiments FCM4, FCM5 we use the notation

\[
\tilde{\eta}_+(\ell\tau) = \max \left\{ \max_{j=1,\ldots,N} \eta(\ell\tau, jh), \frac{2FL}{h\nu^2\kappa_{\text{eff}}RT} I(\ell\tau) \right\} .
\] (4.108)

In Figures 4.7-4.9 the first axis is the spatial variable $x$, the second one is time and the third one is overpotential at the nodes (the numerical solution for the unknown function $\eta$).

First we applied an appropriate time step (see Fig. 4.7, FCM1), i.e., one that can guarantee the non-negativity of the solution (see Theorem 4.4.1). In this case we have a smooth non-negative solution, which was our expectation based on the results of the previous sections. If we applied a time step size which is violating the given condition (4.27), we got unstable results, and after some time step at some nodes the solution goes below zero (see Fig. 4.8, FCM2). However, when we repeated the experiment with $\theta = 0.8$ and we applied an even larger time step (FCM3), we got a stable result, which is non-negative. Since we gave only a sufficient condition in Theorem 4.4.1 for the non-negativity preservation of the method, this experiment also supports our theoretical results.
Table 4.3: The parameters of the macro-homogeneous numerical models FCM1-FCM5, PREC, SEQ12, SEQ21, SYM, SYM2-SYM4 (for FCM3 we applied $\theta = 0.8$ and for the experiments on operator splitting methods $\theta = 0.5$)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h$</td>
<td>0.1</td>
<td>$L$</td>
<td>0.0032</td>
</tr>
<tr>
<td>$N$</td>
<td>10</td>
<td>$\theta$</td>
<td>0.4 (0.8, 0.5)</td>
</tr>
<tr>
<td>$\nu^2$</td>
<td>7.38E01</td>
<td>$\kappa_{\text{eff}}$</td>
<td>0.02</td>
</tr>
<tr>
<td>$T$</td>
<td>293</td>
<td>$\sigma_{\text{eff}}$</td>
<td>1.9</td>
</tr>
</tbody>
</table>

Figure 4.7: Smooth, non-negative numerical solution of the macro-homogeneous model obtained by an appropriate time step size (FCM1)

Experiments FCM4 (see Fig. 4.9 for the numerical solution) and FCM5 are prepared by applying time dependent total current density, i.e., boundary condition. In Fig. 4.10 and 4.11 the horizontal axis is the time step or time level and the vertical is the overpotential. In both cases the solid line denotes the maximum of the overpotential at the given time level, and the dotted line is the estimated maximum based on (4.33).

In Figs. 4.10 and 4.11 and in Table 4.4 one can easily see that the theoretical estimation is valid for these experiments, i.e., on the analyzed domain the discrete numerical values are less than the estimated discrete maximum.

It is worth emphasizing that in experiment FCM5 the discrete maximum estimation of the result in the next time step ($\eta_+ (\ell \tau)$) (see Fig. 4.11) was mainly determined by the second term (which is related to the boundary condition). However, when the total current density falls to 0.5 A from 2 A (at $\ell = 90$), then for some time step the discrete maximum
Figure 4.8: Unstable numerical solution of the macro-homogeneous model obtained by too large time step (FCM2) estimation is determined by the maximum of the numerical solution at the previous time level.

For the numerical experiments of the results related to the operator splitting (PREC, SEQ12, SEQ21, SYM, SYM2-SYM4) the applied material and numerical parameters can be found in Table 4.4 and Table 4.5, where according to conditions (4.42) we denote

\[ \tau_- = \frac{h^2}{2(1 - \theta)} \]  

(4.109)

and \( \tau_{2-} \) denotes the left-hand side of the second condition in (4.42). According to our theoretical results, if the applied time step size is less than \( \tau_{2-} \) in the presented explicit method to obtain the numerical solution of \( P2 \), then the non-negativity of the result is guaranteed.

In what follows PREC denoted the numerical solution of the macro-homogeneous model, which is considered to be the exact solution in the course of the error computation. In its case the time step size was extremely low, and any further decrease affects only the 4th and smaller decimals of the computed results, hence for our purpose it can be taken as the exact solution of the model with the given parameters.

In Fig. 4.14 and Table 4.64 (where \( \epsilon_{\text{max}} \) denotes the error in maximum norm) the numerical error of the presented splitting methods are shown at the time level \( \ell_{\text{max}} \). According to the results, the sequential splitting method SEQ12 has the least error, which can be explained
Figure 4.9: Solution of the macro-homogeneous model obtained by applying an appropriate time step size and time dependent total current density \( I_4(\ell \tau) \) (FCM4).

Figure 4.10: Comparison of the maximal values of the numerical solution and the theoretical estimation of the maximum in \((4.108)\) (FCM4).
Figure 4.11: Comparison of the maximal values of the numerical solution and the theoretical estimation of the maximum in (4.108) (FCM5)

Table 4.4: Time steps of the numerical models, bounds, negative results, extreme values and estimations

<table>
<thead>
<tr>
<th>ID</th>
<th>$\ell_{\text{max}}$</th>
<th>$\tau$</th>
<th>$\tau_2$</th>
<th>$\tau_-$</th>
<th>$\tau_{-2}$</th>
<th>$\eta_-$</th>
<th>$\eta_+$</th>
</tr>
</thead>
<tbody>
<tr>
<td>PREC</td>
<td>6000 1.00E-06</td>
<td>-</td>
<td>8.48E-03</td>
<td>-</td>
<td>0</td>
<td>2.72E-02</td>
<td></td>
</tr>
<tr>
<td>SEQ12</td>
<td>60 1.00E-03</td>
<td>$\tau$</td>
<td>1.23E-02</td>
<td>1.36E-02</td>
<td>0</td>
<td>2.70E-02</td>
<td></td>
</tr>
<tr>
<td>SEQ21</td>
<td>60 1.00E-03</td>
<td>$\tau$</td>
<td>1.23E-02</td>
<td>1.36E-02</td>
<td>0</td>
<td>2.92E-02</td>
<td></td>
</tr>
<tr>
<td>SYM</td>
<td>60 1.00E-03</td>
<td>$\tau$</td>
<td>1.23E-02</td>
<td>1.36E-02</td>
<td>0</td>
<td>2.81E-02</td>
<td></td>
</tr>
<tr>
<td>SYM2</td>
<td>60 2.80E-03</td>
<td>$\tau$</td>
<td>1.23E-02</td>
<td><strong>1.36E-02</strong></td>
<td>-3.99E-03</td>
<td>3.67E-03</td>
<td></td>
</tr>
<tr>
<td>SYM3</td>
<td>60 1.00E-03</td>
<td>$\tau$</td>
<td>1.23E-02</td>
<td>6.78E-03</td>
<td>0</td>
<td>2.24E-02</td>
<td></td>
</tr>
<tr>
<td>SYM4</td>
<td>6 1.00E-02</td>
<td>$\tau$</td>
<td>1.23E-02</td>
<td><strong>6.78E-03</strong></td>
<td>-5.43E-05</td>
<td>3.57E-03</td>
<td></td>
</tr>
<tr>
<td>SYM5</td>
<td>6 1.00E-02</td>
<td>$\tau/10$</td>
<td>1.23E-02</td>
<td>6.78E-03</td>
<td>0</td>
<td>1.49E-02</td>
<td></td>
</tr>
</tbody>
</table>
PROTON EXCHANGE MEMBRANE FUEL CELLS

<table>
<thead>
<tr>
<th>ID</th>
<th>Splitting</th>
<th>Kinetics</th>
<th>$\varepsilon_{\text{max}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>PREC</td>
<td>-</td>
<td>linear</td>
<td>0</td>
</tr>
<tr>
<td>SEQ12</td>
<td>Sequential P1-P2</td>
<td>linear</td>
<td>2.44E-04</td>
</tr>
<tr>
<td>SEQ21</td>
<td>Sequential P2-P1</td>
<td>linear</td>
<td>1.97E-03</td>
</tr>
<tr>
<td>SYM</td>
<td>Symmetric</td>
<td>linear</td>
<td>8.97E-04</td>
</tr>
<tr>
<td>SYM2</td>
<td>Symmetric</td>
<td>linear</td>
<td>2.41E-02</td>
</tr>
<tr>
<td>SYM3</td>
<td>Symmetric</td>
<td>Butler–Volmer</td>
<td>-</td>
</tr>
<tr>
<td>SYM4</td>
<td>Symmetric</td>
<td>Butler–Volmer</td>
<td>-</td>
</tr>
<tr>
<td>SYM5</td>
<td>Symmetric</td>
<td>Butler–Volmer</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 4.5: Splitting methods, applied kinetics and error in maximum norm

by the fact that problem $P2$ gives zero as result for homogeneous initial condition, hence, actually the first step of experiment $SEQ21$ is only a half step. Since the symmetrically weighted splitting method relies on both sequential steps, this error also arises in this case (especially at the boundaries).

In Fig. 4.14 the first figure shows the error in the whole domain, while the second figure presents it on a different scale only at the nodes 6 to 10, in order to make it visible.

For experiments $SYM2$-$SYM5$ we have analyzed the theoretical results of the non-negativity preservation of the symmetrically weighted splitting. In a previous experiment ($SYM$) we applied an appropriate time step size, which can guarantee the non-negativity of the numerical solution (see Fig. 4.12). For $SYM3$ we applied too large time step size, while the rest of the parameters remained the same. The solution (see Fig. 4.13) became unstable with some negative values after a few time steps.

It is worth emphasizing that the two subproblems ($P1$ and $P2$) might have different conditions on the non-negativity preservation of the numerical solution. For the sake of completeness, we made some numerical experiments applying Butler–Volmer kinetics (see (4.10)) to present the case when we apply a time step size which validates our theoretical results for both subproblem ($SYM3$), and a bigger time step size which violates the condition of the non-negativity preservation for the numerical solution $P2$, but not for $P1$ ($SYM4$). In the latter case the results have some negative values (see Table 4.4), however, a great advantage of the operator splitting method can be exploited now, i.e., we can apply two different time step sizes for the two different subproblems to fulfill the given sufficient conditions one by one, in order to get non-negative result ($SYM5$).

For the experiments $FCH1$-$FCH4$ the heterogeneous model was used, which was presented in Section 4.6. In the course of these experiments we applied diffusion kinetics (see (4.11)). The corresponding material and numerical parameters can be found in Table 4.6 and 4.7, where according to the conditions in (4.81) and (4.82) now is $\tau_-$ defined as follows:
Figure 4.12: Smooth, non-negative solution, obtained by an appropriate time step size and symmetrically weighted operator splitting (SYM)

Figure 4.13: Unstable solution with some negative values, obtained by too large time step size and symmetrically weighted operator splitting (SYM2)
PROTON EXCHANGE MEMBRANE FUEL CELLS

Figure 4.14: Absolute error of the presented operator splitting methods (SEQ12, SEQ21, SYM)

<table>
<thead>
<tr>
<th>ID</th>
<th>( \ell_{\text{max}} )</th>
<th>( \tau )</th>
<th>( \tau_- )</th>
<th>( \eta_- )</th>
<th>( I )</th>
<th>Coupled problem</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCH1</td>
<td>1000</td>
<td>1.00E-4</td>
<td>1.84E-04</td>
<td>0</td>
<td>1</td>
<td>No</td>
</tr>
<tr>
<td>FCH2</td>
<td>20</td>
<td>5.00E-3</td>
<td>2.45E-04</td>
<td>-2.39E-01</td>
<td>1</td>
<td>No</td>
</tr>
<tr>
<td>FCH3</td>
<td>100</td>
<td>1.00E-4</td>
<td>4.62E-04</td>
<td>0</td>
<td>1</td>
<td>Yes</td>
</tr>
<tr>
<td>FCH4</td>
<td>100</td>
<td>1.00E-4</td>
<td>4.62E-04</td>
<td>0</td>
<td>2</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Table 4.6: Results and parameters of the heterogeneous model experiments (FCH1-FCH4)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>( N )</td>
<td>50</td>
<td>( a )</td>
<td>1.0E05</td>
</tr>
<tr>
<td>( C_{\text{dl}} )</td>
<td>1.0E-04</td>
<td>( j_d )</td>
<td>1.16E-02</td>
</tr>
<tr>
<td>( T )</td>
<td>293</td>
<td>( L )</td>
<td>1.89E-03</td>
</tr>
<tr>
<td>( \sigma_{\text{eff}} )</td>
<td>1.8</td>
<td>( i_0 )</td>
<td>9.0E-02</td>
</tr>
<tr>
<td>( \kappa_{\text{eff}} [1-10]^* )</td>
<td>2.0E-03</td>
<td>( \kappa_{\text{eff}} [11-50] )</td>
<td>2.2E-03</td>
</tr>
<tr>
<td>( \alpha [1-10] )</td>
<td>0.75</td>
<td>( \alpha [11-50]^* )</td>
<td>1</td>
</tr>
<tr>
<td>( D )</td>
<td>8.72E-4</td>
<td>( D_2 )</td>
<td>6.10E-9</td>
</tr>
<tr>
<td>([O_2]_{L,\text{ref}})</td>
<td>7.20E-5</td>
<td>( L_2 )</td>
<td>1.58E-4</td>
</tr>
</tbody>
</table>

Table 4.7: The parameters of the heterogeneous model experiments (FCH1-FCH4). In the case of \( \kappa_{\text{eff}} \) and \( \alpha \) the bracket means the specified nodes where the given parameters attain the given values. (* For FCH3 and for FCH4 only the marked values were used on the whole domain.)

109
Figure 4.15: Oxygen concentration distribution at two different total current density levels ($FCH3$, $FCH4$)

\[
\tau_- = \min \{K1, K2\},
\]

where

\[
K_1 = \min_{j=2,\ldots,N-1} h \frac{a(jh)C_{dl}(jh)}{\sigma_{eff}(jh) \partial_j hS(jh)}
\]

and

\[
K_2 = \min_{\ell=0,\ldots,\ell_{\text{max}}} \left( \min_{j=1,\ldots,N} \frac{KC_{dl}(jh)u(\ell \tau, jh)}{I(\ell \tau) a(jh) \partial_x S(jh) + i_0(jh) g(u(\ell \tau, jh))} \right).
\]

By the obtained numerical results one can easily see that when we applied a sufficiently small time step size, the results were non-negative in each time step ($FCH1$). However, there is also an example, when too large time step size was applied, and the results became negative at some parts of the domain ($FCH2$).

For the sake of completeness, in the course of experiments $FCH3$ and $FCH4$ we applied the coupled diffusion problem (see Section 4.6.3) to the heterogeneous model. The corresponding material parameters can be found in the second part of Table 4.7. In Fig. 4.15 the oxygen concentration distributions were compared at two different total current density levels. As theoretically expected in the second case, when the load is higher on the cell, the oxygen consumption is also higher, i.e., the concentration is lower on the analyzed domain than in the first case.
Figure 4.16: Results of parameter fitting based on the Levenberg-Marquardt method and the heterogeneous model. The MEA was prepared with 10 percent Nafion® content and it was hot pressed on 80°C and 2 bar.

The model equations contain several material parameters, some of which can be measured directly, while another part of them cannot be measured after preparing the Membrane-Electrode Assembly (MEA) (see Fig. 4.1). In both models there are three unknown parameters, namely, the effective solid phase conductivity ($\kappa_{\text{eff}}$), the exchange current density at the cathode ($i_0$) and the limiting current ($j_{\text{d}}$). There exist formulas for the calculation of these parameters, however, they are rather crude, and they do not describe their dependence on the cell conditions. In our previous work [50], to investigate the real values of the above parameters and their behavior against temperature, pressure and Nafion® content we set the three parameters in such a way that the numerical solution, obtained by the heterogeneous model, is as close to the conducted measurements as possible. For the solution of this problem we used the Levenberg-Marquardt method, which is a clever combination of the well-known Newton and steepest descent methods. (For further details and results see [50]).

For an example of the result of the method see Fig. 4.16 and Table 4.8. In Fig. 4.16 the curves show the cell potential versus the total current density. One can see clearly that the fitted curve is closer to the measurements than the curve obtained by initial guess.

In the course of the fitting we consider the stationary case, i.e., we assume that the quantities in the equation do not depend on time. During the experiments, the cell potential $V$ or $E_{\text{cell}}$, derived from (4.1), is calculated for different constant values of the current density $I$. The three parameters are to be set such that the V-I curve calculated by the model best fits the measured V-I curve.
Table 4.8: Results of parameter fitting based on the Levenberg-Marquardt method and the heterogeneous model after 7 iteration steps. The first column refers to the error according to the measurements.

<table>
<thead>
<tr>
<th>Result</th>
<th>Error in ( | \cdot |_2 )</th>
<th>( \kappa_{\text{eff}} )</th>
<th>( j_d )</th>
<th>( i_0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial</td>
<td>4.60E-02</td>
<td>1.08E-2</td>
<td>8.57E-03</td>
<td>2.50E-08</td>
</tr>
<tr>
<td>Fitted</td>
<td>3.94E-02</td>
<td>7.94E-3</td>
<td>7.93E-03</td>
<td>1.14E-07</td>
</tr>
</tbody>
</table>
Chapter 5

Real-Life Applications of Fuel Cells

5.1 Car-Sharing Service

Nowadays, several hundred millions live in large cities, and almost the same number of people commute between cities and their outskirts every day. Just in Budapest, half a million people commute every day. The present infrastructure (public transport, private cars) is not sufficient, and so more and more people choose alternative forms of traffic. Basically, there are two approaches to attend these people:

1. Public transport systems, where different fixed services cover a given area,

2. Point-to-point type traffic tools, such as cars, bikes, etc.

The throughput of the traffic cannot be further increased by public transport, since it typically requires more than one change for a person to reach his/her destination. Based on some case-study on the public transportation in some cities of the United States, Navin in [83] have established that each minute spent on a bus is associated with 2 min of walk, 5 min of waiting and 7-10 min of waiting before changing a service. Moreover, services must be maintained even when only few people use them; therefore, the efficiency of energy utilization is low and varies within broad bounds. Ball et al. [5] proposed car sharing as a possible option for the solution of the problems of road transportation. It was concluded that it is not enough to take the advanced technology (fuel cells, batteries etc.) into account, but more efficient demand management and an improved driving habit are also important. Very detailed technology analysis has been done with possible end user price of the hydrogen, but no calculation has been performed for a concrete transportation system. No profitability has been considered also by Ajanovic [2], while the possible price of hydrogen from different renewable energy sources available in Austria was compared. The timeframe of the models are not shorter than 50 years, however, investors are interested in much shorter periods of time. Others, like Degiorgis et al. [22], showed a renewable system where hydrogen was produced by electrolysis and used in vehicles.
equipped with internal combustion engines and heat engines. Detailed financial calculation was done, but the economies of scale were not analyzed, neither the effect of scaling up the different system parameters.

In our previous work [68] the technical and economical opportunities of a hydrogen fuel cell based urban mini-car fleet has been analyzed for urban car-sharing application by different parameters like

- the fleet size,
- the rating power of the fuel cell (including the efficiency, which has been analyzed and developed by the help of the mathematical and numerical models of Chapter 4),
- the building cost of the vehicle,
- the cost of the energy and the hydrogen.

Two different operating strategies have been taken into account:

- Hydrogen comes from regular supply at public hydrogen stations.
- Hydrogen is generated by electrolysis and all of the byproducts are utilized (sold) like oxygen and heat.

The detailed research and analysis were partly based on the HY-GO™, i.e., on the first Hungarian fuel cell vehicle (see Fig. 5.1). Our team has built two versions of the car to attend the Széchenyi Race® (Competition of Alternatively Driven Vehicles, Győr, Hungary) in 2009 and 2010. Our vehicles have been awarded twice by the first prize of their category and the prize of the most innovative vehicle of the race too. (It is worth emphasizing that the author of this work was the race car pilot in both cases.)

The power source of the cars were fuel cells, designed by our research group called “fuelcell.hu”, which has been founded to develop cost-effective technical processes for the industry by merging the electrochemical, mathematical and IT knowledge. In the course of the design, the results of the reliable numerical models (see Chapter 4) were utilized, too.

The electrically driven hub motors of the vehicle are built on a three-wheeled undercarriage, and the energy is stored in the form of hydrogen. Its top speed is 50-60 km/h, which is more than sufficient in cities (especially in city centers). It has a 2 kW liquid cooled fuel cell and a 15 Ah battery pack. A maximum of 360 g hydrogen is stored in the metal hydride canisters. The excess heat of the fuel cell warms the metal hydride canisters to reach the maximum amount of hydrogen. By one refueling the vehicle can go
for 5 hours (or 200 km), and it can be refueled in 2 hours. The inner part of the vehicle is comfortable and carefully designed. Due to their top speed and high maneuverability, these cars represent one of the fastest forms of urban traffic.

The major advantage of hydrogen-driven vehicles is the possibility of quick refueling, which could be a critical point in a business plan built on a service where the increase of utilization guarantees the return of the investment. This point is less critical regarding privately owned cars; however, urban traffic cannot be maintained only by private cars. In this manner, the fuel cell based high pressure hydrogen cars are more sufficient for a car-sharing model because the necessary hydrogen refueling time is close to the refueling time of fossil fuels, but the low pressure metal hydride canisters can be more sufficient for the current maturity of the hydrogen infrastructure.

5.2 Cogeneration Power Plant

In 2008 a consortium of two private companies (STS Group Co. Ltd., Mool Invest Ltd.) and two departments of the Eötvös Loránd University, Faculty of Science (Applied Analysis and Computational Mathematics, Laboratory of Electrochemistry and Electroanalytical Chemistry) have been created to start a 3-year project called ReCoMend. The aim of the project is to develop a small co-generation power plant with regenerative hydrogen fuel cells, which could be an environment-friendly alternative to gas motors used at present for the electric and thermal energy supply of public institutions and family homes.
Gas motors have lately been extremely popular among institutions, firms and local governments, due to the possibility of optimizing the energy consumption. Gas motors simultaneously produce utilizable heat and electric energy, that is why they are called co-generation power plants. Their power range is typically 0.5-1 MW, and they are mainly operated as compensation power plants in peak hours, i.e., when electricity is especially expensive. Their advantage is the high (70-80 percent) efficiency of transforming the energy to electricity and heat, but the drawback is the lack of energy storage capability. In this segment, the schedule-following co-generation power plants, i.e., those that follow and span the time differences between consumption and production (the periods of peak and rest hours) result in direct savings. Therefore, the aim of the development is a co-generation, schedule-following power plant for institutional and average consumer groups, in the first step with a power of 100 kW. The use of the fuel cell allows us to temporarily store the energy, increases the efficiency, and provides environment-friendly functioning (see Fig. 5.2 for a fuel cell prototype). Moreover, it decreases the dependence on the increasingly expensive fossil energy sources due to the use of hydrogen fuel.

The international novelty of the project is that a special catalyst layer is going to be developed for the fuel cell by the help of the presented reliable models in Chapter 4. The new catalyst layer allows the cell to be used not only in one direction, i.e., for the
production of current and heat, but also in a reverse manner, when hydrogen and heat are produced from electricity, i.e., energy is stored. The functioning of the cell can be fully regularized, which allows us to follow the schedule, and so to optimize the current uptake on an institutional level and (if this tool becomes widely applied) even for the region of the whole country. The integration of a fuel cell and a water electrolyzer in one nano-layer reduces the costs of the device by 50 percent (since one device serves for two functions), which contributes to the widespread use of the product and makes it competitive on an international level.

The small power plant to be developed in the framework of the ReCoMend project will provide the following benefits for the users:

- Environment-friendly and quiet functioning, as opposed to the gas motors used at present (which are characterized by a rev of 10,000, and make a corresponding noise).

- It provides considerable saving by today’s high prices of the fossil energy sources.

- It transforms and stores the produced energy with a high efficiency (more than 80 percent).

- It spans the temporal difference between the production and use of the (alternative) energy, so the schedule can be followed (back and forth functioning).

- The energy will not be stored in a centralized manner at the power stations, but in a decentralized way (decentralized network) in the form of hydrogen, in high-pressure gas bottles or low-pressure metal-hydrid containers.

- It provides multi-purpose energy use in the form of electric current, heat and hydrogen (fuel).

According to the project "Hungary’s energy policy in 2007-2020", it is a fundamental principle of increasing the renewable energy use in such a way that the proportion of the renewable sources should be increased in accordance with the capabilities and the current strength of the country, while not reducing its economical competitiveness.
Summary

Partial differential equations of elliptic and parabolic types can mathematically describe many important phenomena such as heat conduction or reaction-diffusion. Most of these equations can only be solved by some numerical methods, hence, it is very natural that we want to use such discrete models which preserve certain equivalents of the original qualitative properties of the equations, such as the maximum-minimum principle and its special case, the non-negativity preservation.

Many partial differential equation are able to describe such real-life problems, the solutions of which could directly or indirectly change or improve the life quality of the humanity. In the course of the recent developments of many different devices, the appropriate numerical models have often played key roles. One good example is a very promising solution to the energy problem of the world, namely, the fuel cells. For a better understanding of the operation of these devices the application of different mathematical models are very useful in many situations. The main results of this work are the following.

In Chapter 2 we combine several available theoretical estimates in order to obtain a priori two-sided bounds for the classical solutions of elliptic problems with positive reactive terms for arbitrary source functions, and prove the validity of their discrete analogues for some well-known numerical techniques, e.g., finite difference or finite element methods.

In Chapter 3 we investigate the non-negativity preservation for parabolic problems. First the semidiscrete solutions were analyzed and then the exact conditions for the non-negativity preservation in one dimension for the linear finite element method were given, and the conditions under which the bilinear finite element method is non-negativity preserving in two dimensions were formulated.

In Chapter 4 we present the proton exchange membrane fuel cells and its governing equations which are able to mathematically describe the operation of these devices. Two different approaches were derived, and the preservation of some qualitative properties were analyzed as well.

At the end of each chapter our theoretical results were validated by some numerical experiments. Finally, in the last short chapter two different applications of the proton exchange membrane fuel cells were presented.
Az elliptikus és parabolikus parciális differenciálegyenletek sok olyan fontos jelenséget képesek leírni nagy pontossággal, mint például a hővezetés vagy a reakció-diffúzió folyama. Mivel a legtöbb ilyen egyenletet csak numerikusan lehet megoldani, így alapvető elvárás, hogy az így kapott diszkrét modellek rendelkezzenek a folytonos feladat olyan alapvető kvalitatív tulajdonságával, mint amilyen a maximum-minimum elv vagy a nemnegativitás megőrzése.


A második fejezetben néhány korábbi elméleti eredményt felhasználva pozitív reaktív taggal, tetszőleges forrással ellátott elliptikus feladatok klasszikus megoldására adtunk két oldalú a priori becsüléseket, és olyan jól ismert numerikus módszereket esetében, mint a véges elem ill. véges differencia módszere, bizonyítottuk a diszkrét megfelelőik érvényességét.

A harmadik fejezetben prabolikus feladatokra viszgáltuk a nemnegativitás megőrzését. Először a szemidiszkrit megoldásokat elemeztük, majd később megadtuk a nemnegativitás megőrzésének pontos feltételét egy illetve két dimenzióban lineáris és bilineáris véges elem módszer esetén.

A negyedik fejezetben bemutattuk a protoncsere membrános üzemanyagcellákat, illetve azokat a kormányzó egyenleteket, amelyek jól modellezzük a működésüket. Két különböző megközelítést is alkalmaztunk, és a korábbi fejezetekben bemutatott kvalitatív tulajdonságok megőrzésének feltételeit is megvizsgáltuk.

Az egyes fejezetek végén az elméleti eredményeket numerikus kísérletekkel is alátámasztottuk. Végül a legutolsó rövid részben a protoncsere membrános üzemanyagcellák felhasználására mutattuk két gyakorlati példát.
Bibliography


